Gratings: Theory and Numeric Applications, Second Revisited Edition

Tryfon Antonakakis, Fadi Issam Baida, Abderrahmane Belkhir, Kirill Cherednichenko, Shane Cooper, Richard Craster, Guillaume Demésy, John Desanto,, Gérard Granet, Boris Gralak, et al.

Gratings:
Theory and Numeric Applications
Second Revisited Edition

Tryfon Antonakakis
Fadi Baïda
Abderrahmane Belkhir
Kirill Cherednichenko
Shane Cooper
Richard Craster
Guillaume Demesy
John DeSanto
Gérard Granet
Boris Gralak
Leonid Goray
Sébastien Guenneau
Lifeng Li
Daniel Maystre
André Nicolet
Evgeny Popov
Gunther Schmidt
Elizabeth Skelton
Brian Stout
Fréderic Zolla
Benjamin Vial

Evgeny Popov, Editor

Institut Fresnel, Université d'Aix-Marseille, Marseille, France
Institut FEMTO-ST, Université de Franche-Comté, Besançon, France
Institut Pascal, Université Blaise Pascal, Clermont-Ferrand, France
Colorado School of Mines, Golden, USA
CERN, Geneva, Switzerland
Imperial College London, UK
Cardiff University, Cardiff, UK
Université Mouloud Mammeri, Tizi-Ouzou, Algeria
Saint Petersburg Academic University, Saint Petersburg, Russian Federation
Institute for Analytical Instrumentation, Saint Petersburg, Russian Federation
Weierstrass Institute of Applied Analysis and Stochastics, Berlin, Germany
Tsinghua University, Beijing, China

ISBN: 978-2-85399-943-4
www.fresnel.fr/numerical-grating-book-2

ISBN: 2-85399-943-4

Second Edition, 2014

World Wide Web:

www.fresnel.fr/numerical-grating-book-2

Aix Marseille Université, CNRS, Centrale Marseille, Institut Fresnel UMR 7249,
13397 Marseille,
France

Gratings: Theory and Numeric Applications, Second Revisited Edition, Evgeny Popov, editor (Aix Marseille Université, CNRS, Centrale Marseille, Institut Fresnel UMR 7249, 2014)

2014
Presses Universitaires de Provence

Preface to the Second Edition

Evgeny Popov

Aix-Marseille Université, CNRS Central Marseille, Institut Fresnel UMR 7249
Campus de Saint Jerome,13013 Marseille, France
e.popov@fresnel.fr *www.fresnel.fr/perso/popov*

One and a half year ago we have assembled the online edition of our e-book devoted to the theory of diffraction grating. More than a thousand downloads have been registered since then, and I sincerely hope that the detailed description that we tried to present in the book would be useful to the scientific community.

There is no substantial advance in the grating theory during the last 18 months. However, due to time constraints, in the first edition we were not able to include some important developments that have been advanced before. This is the main reason to propose a Second Edition of "Gratings: Theory and Numeric Applications," that contains two more chapters and supplements to other four chapters.

Here is the summary of the changes that are made:

1. Lifeng Li has written Chapter 13 on Fourier Modal Theory, also known in the literature as Rigorous Coupled Wave (RCW) method, a method that proved itself quite efficient for vertically invariant periodic structures (e.g., lamellar gratings). Li's contributions to the method have also given the possibility to improve the differential method applied to arbitrary shaped profile (Chapter 7).
2. Another new contribution is made by L. Goray and G. Schmidt (Chapter 12) concerning the Boundary Integral Method. In the first edition, we have already included a chapter (Chapter 4) on the Integral method. The new contribution develops further the formulation to conical diffraction, and gives a detailed description of the theory, which is the base of two commercially available numerical codes. I hope that this chapter would be useful to the users of these packages.
3. The revised Chapter 6 includes more details on the T-matrix method of section 6.3. Discussion of quasi-modes is also brought more up to date in light of ongoing developments. The notation are slightly modified in the Lattice sum section 6.6 to facilitate complex frequency approaches to quasi-modes, and of few of the lattice sum expressions of this chapter are replaced by slightly simpler expression.
4. Chapter 10 that describes the so called Exact Modal Methods has been completed with the extension of the method to the case of lamellar gratings made of infinitely conducting metal that cannot be treated by the Fourier modal method or the differential method.

5. Chapter 11 devoted to the homogenization techniques is extended to higher-frequency electromagnetic fields that apply for diffraction gratings, treated in a new section 11.4.2
6. I have introduced several minor changes in the description of the differential method (Chapter 7). First, the title is changed by replacing “theory” by “method” in order to be consistent with the other chapters. Second, a new section is included (sec. 7.7 in the Second Edition), which describes how one can avoid discrete Fast Fourier transform (FFT) for gratings with two-dimensional periodicity, by making analytical Fourier transform for soma specific profiles.

Marseille, France
April 2014

Editorial Preface

A typical question that almost all of us (the authors' team and other colleagues) has been asked not only once has in general the meaning (although usually being shorter): "What is the best method for modeling of light diffraction by periodic structures?" Unfortunately for the grating codes users, and quite fortunately for the theoreticians and code developers, the answer is quite short, there is no such a bird like the best method.

In the more than 30 years active studies on the subject, I have worked on the theory and numerical applications of several approximate methods, like Rayleigh expansion, coupled-wave theory, beam propagation method, first-order approximations, singular Green's function approximation, effective index medium theory, etc. My conviction is that they are quite useful (otherwise why to exist) for physical understanding, but my heart lies in what is considered as rigorous grating theories. Name 'rigorous' is used in the sense that in establishing the theories, exact vector macroscopic Maxwell equations and boundary conditions are applied without approximations. The approaches become approximate after computer implementation, due to the impossibility to work with infinite number of equations and unknowns, and due to the finite length of the computer word.

Of course, there are always initial approximations and assumptions, like the infinite dimensions of the grating plane, linearity of the optical response, etc. From physical point of view, the main feature of the methods, presented in this book are characterized by the use of optical parameters of different substances as something given by other physical optics theories and the experiment as an ultimate judge.

The necessity to use more than a single rigorous method comes from practice: different optogeometrical structures made of different materials and working in different spectral regions require a variety of methods, because each one is more effective in some cases, and less effective (or failing completely) in others. In addition, each approach is a subject of constant research and development. Grating modeling, grating manufacturing and grating use go hand in hand, and practice provides strong stimuli for the theory development. Vice versa, recent grating technologies and application cannot advance without proper theoretical and numerical support.

When I started my grating studies, the method of coordinate transformations that uses eigenvector technique to integrate the Maxwell equations (sometimes known as the C-method) has just been formulated. It worked perfectly for holographic grating whatever the polarization and the grating material, but failed completely for grooves with steep facets. It took more than 15 years to refine its formulation, so that now it can deal with echelles and pyramidal bumps (in the case of two-dimensional periodicity) with slopes up to 87 deg steepness. However, the method is not at all adapted to lamellar gratings. On the other hand, the Fourier modal method (also known as Rigorous coupled-wave approach, RCW) is perfect for such profiles, but its use in the case of arbitrary grating profiles (e.g., sinusoidal or triangular profiles) in case of metallic grating material causes problems when using a staircase approximation of the profile. The differential method does not use this approximation and could deal with arbitrary profiles, but it took more than 20 years to make it working with

metallic gratings in TM (p, or S) polarization. And quite ironically, the improvement came from advances in the competing RCW approach.

These methods are relatively easy for programing nowadays, after solving the numerical problems due to growing exponentials and factorization rules of the product of permittivity and electric field, however there are still some persisting problems for highly conduction metals. In addition, neither the differential, nor the Fourier modal methods can deal with infinitely conducting gratings.

Several methods are quite flexible concerning the geometry of the diffracting objects and the grating material. For example, the integral method can treat inverted profiles, rod gratings with arbitrary cross section, finitely or infinitely conducting materials in any polarization, but its programming require deep mathematical understanding of the singularities and integrability of the Green's functions. Other two flexible methods are quite famous and widely used, even in the form of commercially available codes. These are the finite-element method, and the finite-difference time domain method. The flexibility with respect to the geometrical structure, optical index inhomogeneity and anisotropy, etc. has to be paid by the necessity of sophisticated meshing algorithms and very large sparse matrix manipulations.

These few examples represent only the top of the iceberg, and are invoked to illustrate the basic idea that *the best method has not been invented, yet*. Probably never.

We have tried to gather a team of specialists in rigorous theories of gratings in order to cover as large variety of methods and applications as practically possible. The last such effort dates quite long ago, and it has resulted in the famous *Electromagnetic Theory of Gratings* (ed. R. Petit, Springer, 1980), a book that has long served the community of researchers and optical engineers, but that is now out of press and requires a lot of update and upgrade, something that we hope to achieve, at least partially with this new book.

Our choice of electronic publishing is determined by the desire to ensure larger free access that is not easily available through printed editions. I want to thank all the contributors to this Edition. Special thanks are due to my colleagues Fréderic Forestier and Boris Gralak for the technical efforts to make the electronic publishing possible.

Marseille, France Evgeny Popov
December 2012

Chapter 1:

Introduction to Diffraction Gratings: Summary of Applications

Evgeny Popov

Table of Contents:

Chapter 1

Introduction to Diffraction Gratings: Summary of Applications

Evgeny Popov

Aix-Marseille Université, CNRS, Centrale Marseille, Institut Fresnel UMR 7249,
Campus de Saint Jerome,13013 Marseille, France
e.popov@fresnel.fr www.fresnel.fr/perso/popov

Periodic systems play an important role in science and technology. Moreover, desire for order in Nature and human society has accompanied development of philosophy. Simple periodical oscillation is referenced as 'harmonic' in mechanics, optics, music, etc., the name deriving from the Greek *ἁρμονία* (harmonía), meaning "joint, agreement, concord" [1.1]. It is not our purpose here to study harmony in general, nor harmony in physics. We are aiming to much more modest target: rigorous methods of modeling light propagation and diffraction by periodic media.

The methods presented in the book have already shown their validity and use from x-ray domain to MW region, for nonmagnetic and magnetic materials, metals and dielectrics, linear and nonlinear optical effects. The existing variety of these methods is due not only to historical reasons, but mainly to the absence of The Method, a universal approach that could solve all diffraction problems. Some of the approaches cover greater domain of problems, but more specialized ones are generally more efficient. The other reason of the great number of methods is the complexity and variety of their objects and applications.

1.1. Diffraction property of periodic media

The most important property of diffraction grating to create diffraction orders has been documented by Rittenhause for the first time in 1786 [1.2] due to the observation made by Francis Hopkinson through a silk handkerchief. The appearance of diffraction orders rather than the specularly reflected and transmitted beams was studied experimentally by Young in 1803 [1.3] with his discovery of the sine rule. A detailed presentation of the analytic properties of gratings can be found in Chapter 2.

Secondary-school pupils are supposed nowadays to know the Snell-Descartes law, which undergraduate students in universities are supposed to be able to demonstrate: single-ray diffraction on a plane interface results in a single transmitted and single reflected rays. The critical advantage of periodic perturbation of the interface (variation of the refractive index or surface corrugation) changes the impulsion (wavevector surface component $\vec{k}_{S,m}$) of the incident wave $\vec{k}_{S,i}$ along the surface by adding or subtracting an integer number of grating impulses (grating vectors) $\vec{K}$:

$$\vec{k}_{S,m} = \vec{k}_{S,i} + m\vec{K} \tag{1.1}$$

where $\vec{K} = \frac{2\pi}{d}\hat{d}$ and d is the grating period in a unit-vector direction $\hat{d}$.

If the interface lies in the xy-plane and the periodicity is along the x-axis (Fig.1.1), and the incidence lies in a plane perpendicular to the grooves, the equation in reflection takes the form of the so-called *grating equation*:

$$\sin\theta_m = \sin\theta_i + m\frac{\lambda}{d} \tag{1.2}$$

where λ is the wavelength of light.

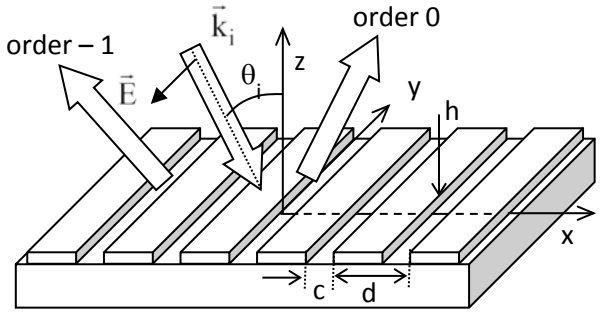


Fig.1.1. Lamellar grating working in in-plane regime in TM (transverse magnetic) polarization, together with the coordinate system, incident wavevector.

The case of conical (off-plane) diffraction by a plane grating having one-dimensional periodicity can also be described by eq.(1.1), preserving the wavevector component parallel to the groove direction:

$$\begin{aligned} k_{y,m} &= k_{y,i} \\ k_{x,m} &= k_{x,i} + mK \\ k_{z,m} &= \sqrt{k^2 - k_{x,m}^2 - k_{y,i}^2} \end{aligned} \tag{1.3}$$

where k is the wavenumber in the cladding. As a result, the diffracted beams lie on a cone, thus the name *conical diffraction*.

Two-dimensional periodicity (having grating vectors $\vec{K}_1$ and $\vec{K}_2$) imposed on the optogeometrical properties of the plane interface creates two sets of diffraction orders together with their spatial combinations, subjected to the same rule formulated just before eq.(1.1):

$$\vec{k}_{S,mn} = \vec{k}_{S,i} + m\vec{K}_1 + n\vec{K}_2 \tag{1.4}$$

Pure three-dimensional (3D) periodicity appears in crystallography and photonic crystals, if the substance is assumed to fill the entire space. The third-direction periodicity imposes additional condition to the wavenumber (said more precisely, to $k_{z,m}$, which is already defined by the wave equation as given in the third equation of (1.3)), which leads to creation of discrete modes propagating in 3D periodic structures. This also leads to the appearance of propagating and forbidden zones structure.

1.2. Classical gratings in spectroscopy

The most common application of diffraction gratings is due to the fact that outside of the specular order (m = 0), the diffraction order direction depends on the wavelength, as stated in eq.(1.2). The result is that the grating acts as a dispersive optical component, with several advantages when compared to the prisms:

1. The grating can be a plane device, while the prism is a bulk one that requires larger volumes of optically pure glass (to add the difficulties of weight and temperature expansion constrains).
2. Provided a suitable reflecting material, the grating can work in spectral regions, where there is no transparent 'glass' with sufficient dispersion.
3. Grating dispersion can be varied, as it depends on the groove period, while prism dispersion depends on the material choice and groove angle, which gives quite limited choices.

Since the first works of Young, followed by Fraunhofer's quite serious attention to diffraction gratings use and properties [1.4], there is rarely more important device in spectroscopy achievements that lay the basis of modern physics. An interested reader can find some important aspects of their history, properties, and application in [1.5]. Despite the above-listed advantages compared to prisms, the diffraction gratings never have sufficient performance for their spectroscopy customers:

1. *There is no enough diffraction efficiency*, defined as the ratio of the incident light diffracted in the order used by the application.

This is probably the problem that has been mostly treated by rigorous grating methods, because they are the only ones to provide feasible results on the energy distribution, while the other characteristics (spectral resolution, scatter, dispersion, order overlap, etc.) can be obtained by simpler approaches.

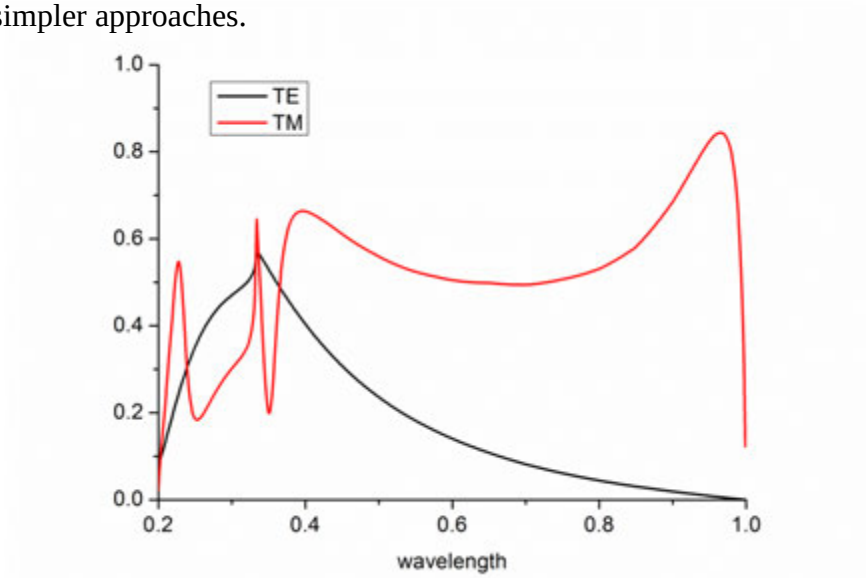


Fig.1.2. Diffraction efficiency of aluminum-made sinusoidal grating with period d = 0.5 μm as a function of the wavelength. Littrow mount (–1st order diffracted in the direction of the incident beam) for the two fundamental polarizations (transverse electric TE and transverse magnetic TM).

A typical spectral dependence of the diffraction efficiency (defined more precisely as the ratio between the energy flow in the corresponding order and the energy flow in the incident order in z-direction) of a surface-relief sinusoidal grating made of aluminum and working in the -1^{st} Littrow mount[1] is presented in Fig.1.2. As observed, the problems of spectral variation of efficiency adds to its polarization dependence.

2. *There is no enough resolution,* defined as the ratio between the working wavelength and the smallest distinguishable (as usual, defined using the Rayleigh criterion) spectral interval:

[1] Littrow mount means retrodiffusion, when the diffracted beams propagates in a direction opposite to the incident beam, i.e. $\sin\theta_{-1} = -\sin\theta_i$

$$R = \frac{\lambda}{\Delta\lambda} \tag{1.5}$$

Classical diffraction theories show that the spectral resolution is proportional to the number of grooves illuminated by the incident beam, and inversely proportional to $\cos(\theta_i)$, if we consider 1D periodicity in non-conical diffraction (speaking more precisely, the inverse proportionality implies to the sum of the cosines of the incident and the diffracted angles). The latter dependence gives advantages to grazing incidence for higher spectral resolution applications, namely astronomy.

3. *The efficiency depends on the polarization.*

Laser resonators usually use Brewster windows, so that the requirements are for high-efficient grating working in TM (transverse magnetic) polarization. However, spectroscopic applications do not like this at all. In astronomy, this property can be quite costly, because the loss of a half of the incident light intensity requires twice the exposer time. Low polarization dependent losses (PDL) are one of the most important criterions in optical communications, in general, and in grating applications for wavelength demultiplexing, in particular, necessary for multichannel optical connections. Fortunately, contrary to stellar spectroscopy, gratings used in optical communications work in only very small spectral interval, and are used in quite smaller sizes, so that there exist several solutions that provide high efficiency in unpolarized light over a limited spectral region. The idea is to shift the maxima in the spectral dependence of the two polarizations in Fig.1.2 in order to make them overlap at some required wavelength. One solution is to use a grating having two-dimensional (2D) periodicity, as shown further on in Fig.7.1. The period in the perpendicular direction is sufficiently small as not to introduce additional diffraction orders, and this additional corrugation can shift the position of the TE maximum to longer wavelengths [1.6].

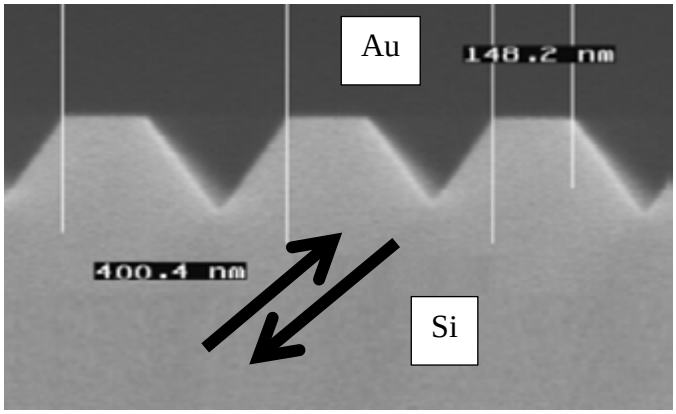


Fig.1.3. SEM picture of the profile of a grating etched in Si wafer and used for wavelength demutiplexing (after [1.7], with the publisher's permission)

Another more conventional solution is to use a classical grating with a 1D periodicity, but having a deformed profile in order to perform the same shift of the TE maximum [1.7, 8]. This can be achieved by introducing a flat region at the bottom of the grooves and "sharpening" the groove triangle, which needs a sharper apex angle. While this is quite difficult to be made with grating ruling or holographic recording, etching in crystalline silicon naturally produces grooves with 70.5° apex angle, as observed in Fig.1.3. A flat region on the top is made if the etching is not complete. When covered with gold, such grating can be used from the silicon side, which is transparent at wavelengths around 1.55 μm. Unpolarized

efficiency greater than 80% can be kept over the communication interval of 50-60 nm, as observed in Fig.1.4.

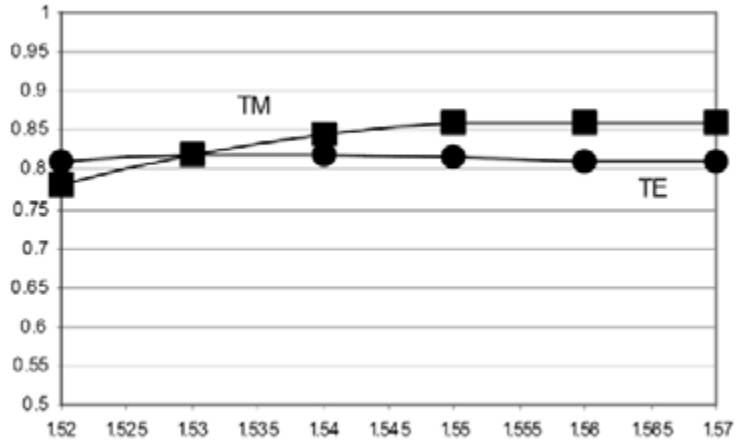


Fig.1.4. Spectral dependence of efficiency in -1st order for the grating shown in Fig.1.3 (after [1.7] , with the publisher's permission).

4. There is an overlap of diffraction orders.

As stated by the grating equation, if the period is chosen to provide diffraction orders inside a large spectral interval, the second diffracted order has the same direction of propagation as the first one for wavelength of light twice shorter. This problem is avoided in many spectroscopic devices by three different methods: additional spectral filtering of undesired orders; interchange of grating having different spatial frequency (period); adding cross-dispersion grating for echelle applications.

In addition, there is a contradiction between some requirements as, for example, free spectral range (spectrum covered by the grating), dispersion and overlap of orders. In some cases, the only compromise is to use interchangeable gratings in order to cover larger spectral range without order overlap and maintaining higher dispersion. Some spectrographs are designed having multiple channels with splitting of the incident beam between them, however reducing the illumination power in each channel.

1.3. Echelle gratings in astronomy

In astronomy, measurements of very low signals coming from far cosmic objects require large periods of time, so that the loss of energy due to low efficiency or/and strong polarization dependence leads to a further growth of costs. When efficiency constraints are added to the independence of the polarization, the only known solution is the echelle grating (high groove-angle triangular groove profile used in grazing incidence, Fig.1.5) that has the advantages of almost equal and high efficiency in both fundamental polarizations [1.9].

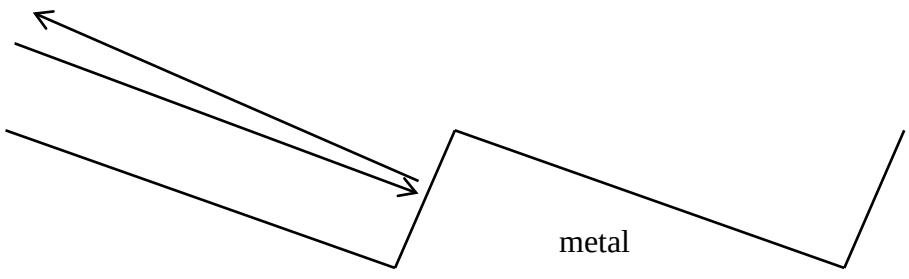


Fig.1.5. Schematical presentation of echelle grating.

The high efficiency in unpolarized light is obtained when the diffraction is made as if light is reflected by the working facet in its normal direction. Of course, it is necessary that the reflections at consecutive facets are in phase. The problem is that the efficiency varies rapidly with the wavelength, and the maxima switch between consecutive orders (Fig.1.6). The separation of orders is usually made using another shorter-period grating with grooves perpendicular to the echelle (called cross-dispersion), so that the different orders are separated in direction perpendicular to the echelle dispersion direction.

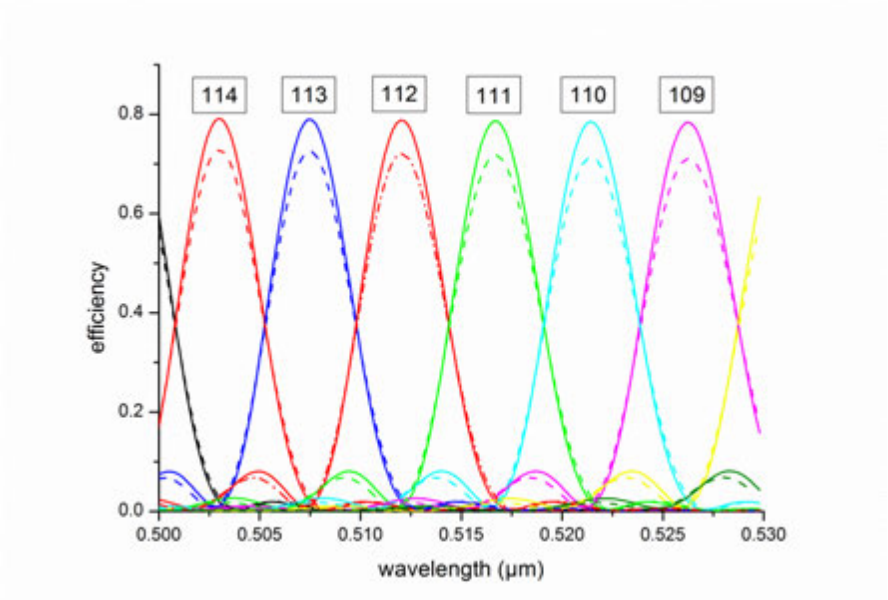


Fig.1.6. Diffraction efficiency of an echelle made of aluminum with 31.6 grooves/mm and 64° groove angle of the working facet. Incident angle is equal to 64°. The numbers of the diffraction orders are indicated in the figure.

Echelle gratings are also used in transmission, glued to the hypotenuse face of a prism that has a role to deviate back the beam diffracted by the grating, so that the principal diffracted order of the device propagates almost in the same direction as the incident beam for a chosen central wavelength. The device is known as a Carpenter prism or GRISM and gives the possibility to convert an imaging device (camera) into a long slit spectrograph [1.10], commonly used in airborne or space borne scientific missions.

UV excimer lasers used in photolithography at 193.3 nm wavelength can be equipped with an echelle grating for narrowing the spectral line. While in lasers working in the visible and IR, the resonators are equipped with diffraction gratings with symmetrical grooves working in -1^{st} order than easily can be made holographically or lithographically, a grating working in the lowest order at 193.3 nm must have more than 9 000 gr/mm, very difficult for fabrication and impossible for replication. Echelles take longer to be made and are more expensive as they require mechanical ruling engines with temperature control and clean environment, but have large periods and can be replicated from the ruled masters and submasters to become available at acceptable prices.

1.4. Gratings as optical filters

Spectroscopic applications of gratings use one or more diffraction orders that differ from the specular reflected and transmitted ones, because of the required spectral dependence. There exist, however, several applications that use the zero order(s), even with corrugation of the

surface or modulation of the refractive index in the grating region. These are devices having refractive or reflection properties in the zeroth order that are modified by the grating surface.

1.4.1. Zero-order diffraction (ZOD) imaging

One of the applications uses the diffraction in higher orders to change the spectral dependency in the zero order, as light that is diffracted in the higher order(s) is absent in the zero order. The structure known as zero-order diffraction *(ZOD)* microimage [1.11] represents a transmission grating (usually with a rectangular or triangular groove form that can be replicated by relief printing in plastic sheet. Appropriately choosing the groove depth, one obtains broad-band color filters in transmission by using non-absorbing materials without colorants that can bleach.

Another type of gratings have periods, smaller than the wavelength in the substrate and the cladding, chosen to avoid the propagation of other that the zero orders. Such structures are known as *subwavelength gratings*, and they play important role in integrated optical devices and recently in plasmonics to transfer energy from one to another guided mode in dielectric or metallic waveguides, or to change the mode direction, or to focus guided light (see Section 1.5).

1.4.2. Surface/guided mode excitation

A subwavelength grating can serve to couple the incident light to surface or cavity resonances that can exist in the grating structure. The absence of higher orders means only that they are evanescent rather than propagating in the cladding or in the substrate. The horizontal component of their propagation constant (say, k_x) is larger than the wavenumber in the surrounding media, and thus it can excite a surface or waveguide mode that is subjected to the same requirement in order to stay confined to the surface. The grating action is the same as in the coupling between the incident wave and one of the diffraction orders, only that now $k_{x,-1}$ is equal to the real part of the mode propagation constant k_g:

$$\mathrm{Re}(k_g) = k_{x,i} - K \tag{1.6}$$

To fail to see the quite small difference between k_g and k_0 in the case of surface plasmons on highly conducting metal surfaces is one of the rare failures of Lord Rayleigh [1.12] when trying to explain Wood's anomalies [1.13]. The case when $k_{x,m} = k_0$ represents a transition of the m-th diffracted order between a propagating and an evanescent type, as follows from eq.(1.3). This transition leads to a redistribution of energy between the propagating orders, observed in efficiency behavior as a phenomenon called cut-off anomaly.

Lord Rayleigh attributed Wood's anomalies to the cut-off of higher orders, whereas the true explanation is that Wood anomaly is due to surface plasmon excitation. It is easy to judge nowadays, but the difference in the spectral and angular positions of the cut-off and surface plasmon anomaly sometimes is smaller than the experimental error, or more important, the error in the knowledge of the grating period. The resonant nature of the surface-plasmon anomaly has much more pronounced features than the cut-off anomaly, a fact first noticed by Fano [1.14] and further developed by Hessel and Oliner [1.15]. The form of the so-called Fano-type anomaly can easily be derived from the interference with a non-resonant contribution (for example, non-resonant reflection by the grating layer) and a resonant effect (for example, excitation by the incident wave of a surface mode that is diffracted back into the direction of the non-resonant wave), as sketched in Fig.1.7.

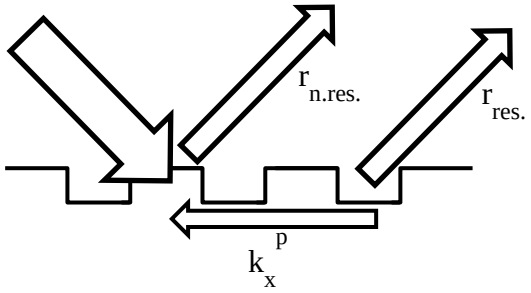


Fig.1.7. Process of interference between the non-resonant and the resonant reflections (second term in eq.(1.7)) that is created due to the excitation of the surface or guided wave through the –1st grating order, and then radiated into the cladding through the +1st diffraction order.

Here, by mode, we mean a surface or guided wave that represents an eigen (proper) solution of the homogeneous diffraction problem (assuming non-zero diffracted field without incident wave). The latter effect is commonly described in physics as having a Lorentzian character (e.g. electric circuit dipole oscillator). For example, the wavelength dependence of the total reflectivity r will be the sum of the non-resonant and the resonant Lorentzian contributions:

$$r = r_{n.res.} + \frac{c_{res.}}{k_{x,i} - k_x^p} \tag{1.7}$$

where the coefficients $r_{n.res.}$ and $c_{res.}$ are slowly varying functions of the incident wave parameters. The pole k_x^p of the resonant term is equal to the surface/guided wave propagation constant. The Fano-type equation is obtained by taking the common denominator in eq.(1.7):

$$r = r_{n.res.} \frac{k_{x,i} - k_x^z}{k_{x,i} - k_x^p} \tag{1.8}$$

with $k_x^z = k_x^p - c_{res.}/r_{n.res.}$ representing a zero of the reflectivity. The formula implies that for each resonance, there exist an associated zero, that is expressed as a minimum of the anomaly. In reality, the pole has always non-zero imaginary part due to the interaction between the incident and the surface/guided wave. The zero takes, in general, complex values, but there are several important practically cases when the zero could become (and remains) real, i.e., the reflectivity can become zero. The imaginary part of the pole determines the quality factor (width) of the resonant maximum.

The same reasoning applies in transmission (and in any other existing propagating order), with the same pole (same resonance), but different zeros. Depending on the imaginary parts of the pole and the zero, sometimes the resonant anomaly can show itself as a Lorentzian maximum, sometimes as a pure minimum on otherwise highly-reflecting background, sometimes both maximum and minimum can manifest themselves. The important cases when the zeros are almost real are used in surface plasmon absorption detectors and in resonant dielectric grating filters (see the next two subsections).

1.4.3. Surface plasmon absorption detector

Wood anomaly can sometimes lead to a total absorption of light by shallow metallic gratings with groove depth that does not exceed 10% of the wavelength. This phenomenon was called *Brewster effect* in metallic gratings [1.16] and has found an important application in chemical

and biochemical surface-plasmon grating detectors [1.17]. The effect of total light absorption appears in a narrow spectral and angular interval (Fig.1.8).

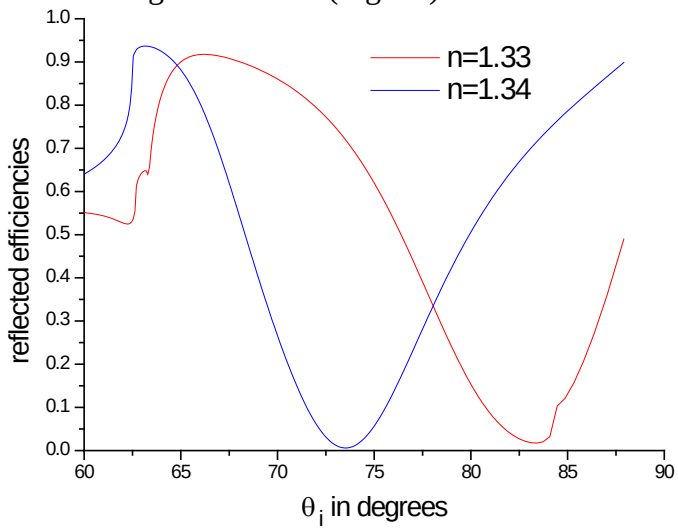


Fig.1.8. Reflectivity of an optical surface plasmon detector as a function of the incident angle. Wavelength equals 850 nm, TM polarization. Cladding is glass, the substrate index for the two curves is indicated in the figure, and the grating layer is made of silver with thickness of 40 nm. The grating profile function contains two Fourier harmonics: $35\sin\left(2\pi x/d\right)+59\sin\left(4\pi x/d+\pi/2\right)$ *with the depths given in nanometers (after [1.18], with the permission of the publisher).*

The position of the anomaly depends on the propagation constant of the plasmon surface wave that is quite sensible to the variation of the refractive index of the surrounding dielectric, thus the determination of the position of the anomaly brings information about the composition of the surrounding medium, i.e., serves as an optical detector (Fig.1.8). Introduction of the second Fourier component of the profile function increases the direct interaction between the plasmon surface waves propagating in opposite directions (the grating vector of the second harmonic is twice longer that of the first one), interaction that increases the sensitivity of the device.

1.4.4. Resonant dielectric filters

There is a particular case when the maximum in the reflectivity due to the resonant waveguide mode excitation stays theoretically at 100% on a low-reflective non-resonant background value. This is the case of corrugated dielectric waveguides having symmetrical grooves. If, in addition, the substrate is identical to the cladding, the zero in reflection remains real whatever the other parameters, i.e. a 100% maximum is accompanied by a 0 value minimum in the reflectivity (Fig.1.9). This peculiarity was accidentally found in 1983 [1.19, 20] followed by a theoretical explication in 1984 [1.21]. The effect has been independently rediscovered in 1989 by Magnusson [1.22], who has done a lot for its analysis, and quite important practically, for its extension in transmission [1.23]. The advantage of the device is that the quality factor increases with the decrease of the groove depth, i.e. very narrow-line spectral filters can be obtained using shallow gratings, at least theoretically. However, as usual, the advantage is paid back somewhere, namely in the tight angular tolerances: when eq.(1.6) results in high spectral sensibility, it also is responsible for strong angular sensibility. Sentenac and Fehrembach [1.24] proposed to introduce a direct coupling between the waveguide modes propagating in the opposite direction (as was done in Fig.1.8), but this time aiming to a significant *reduction* of the angular sensibility of the effect.

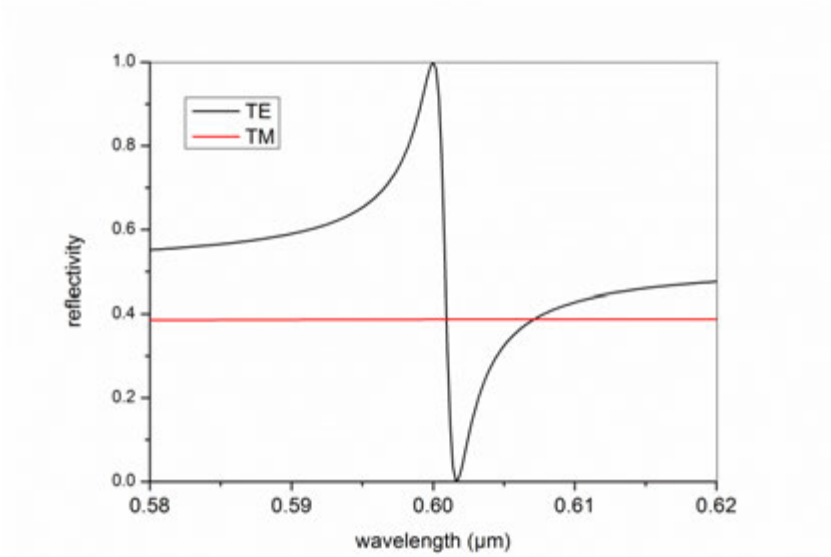


Fig.1.9. Reflection by a corrugated dielectric waveguide. Symmetrical triangular profile with groove angle equal to 10°. Substrate and cladding have optical index equal to1, the layer has index 2.3 and its thickness is equal to 69 nm. Incident angle is equal to 26.7°. Excitation of TE waveguide mode leads to a narrow spectral anomaly.

The idea, was based on the flattening of the angular dependence of the guided wave propagation constant on the boundaries of the forbidden gaps, created by the direct mode coupling due to the structure periodicity, one of the basic properties of photonic (and electronic) crystals.

1.4.5. Enhanced transmission through hole arrays in metallic screens

In 1998 Ebbesen et al. [1.25] reported an interesting effect on metallic screen perforated with circular holes (Fig.1.10a). They were surprised to observe peaks with relatively strong transmission in the spectral dependence (Fig.1.11).

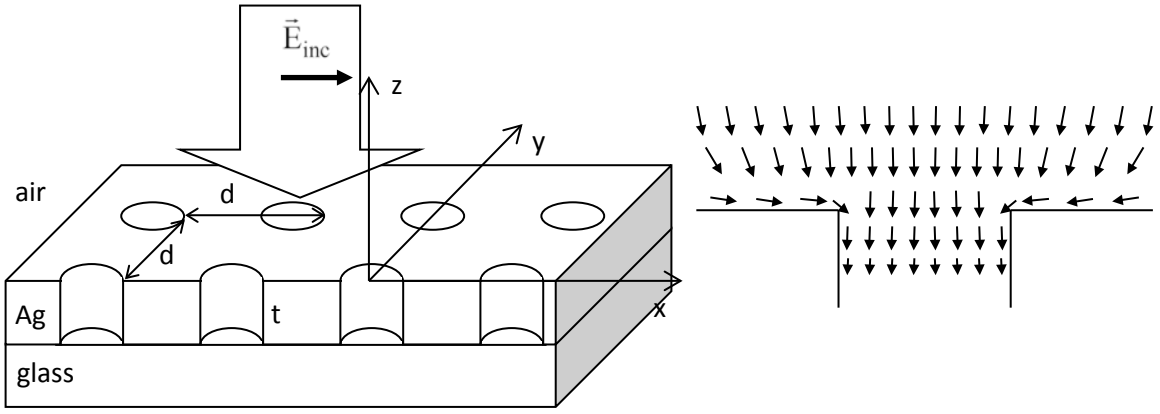


Fig.1.10. (a) Schematical representation and notations of a two-dimensional hole array perforated in a metallic screen deposited on a glass substrate and illuminated from above with linearly polarized incident wave. (b) Surface-plasmon assisted energy flow inside the aperture close to the resonance.

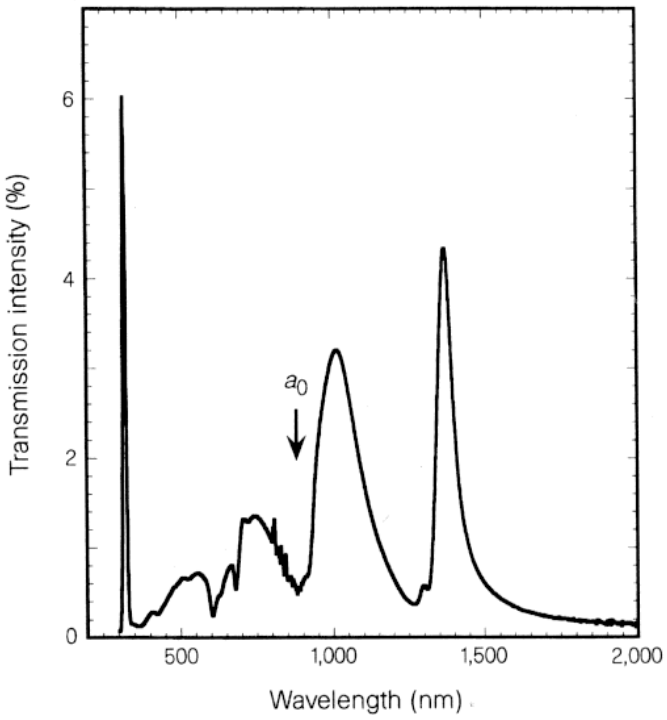


Fig.1.11. Spectral dependence of the transmission of the structure presented in Fig.1.10a, with d = 0.9 μm, t = 0.2 μm, and hole diameter of 0.2 μm (after [1.25], with the publisher's permission).

These peaks were identified as resulting from surface plasmon excitation on the upper, or the lower surface. Many theoretical and experimental works appeared as a result of this discovery, which revived the interest to surface plasmons and grating structure, leading to a new name of the plasmon studies called now *plasmonics*.

It seems nowadays that the explanation of the enhanced transmission observed in Fig.1.11 lies in the common action of two phenomena [1.26]. The first one is the surface plasmon excitation that enhances the electromagnetic field intensity near the entrance apertures, as represented schematically in Fig.1.10b. The second phenomenon is the tunneling of the mode in the vertical hollow waveguides inside the holes. Although for small holes even the lowest mode is evanescent (contrary to 1D lamellar gratings, where TEM mode exists without cut-off), it gives the possibility of energy transfer from the upper to the lower interface, much more efficiently than the tunneling through the non-perforated screen [1.27].

One unexpected consequence of the observation of Ebbesen et al. came in fluorescence single-molecule microscopy and in the biomembrane studies. Both require small measuring volumes in order to study only a very small number of molecules or a small portion of the membrane surface. However, small measuring volumes mean week signals. Here comes the role of the surface plasmon enhanced field and evanescent mode inside the aperture. When the mode is close to its cut-off, the real part of its propagation constant along the axis of the hole tends to zero, which leads to a compression of the electromagnetic field at the entrance aperture, increasing even more the field density and thus the measured signal. In addition, as the field is evanescent inside the hole, the effective measuring volume does not extend to the entire hole depth (see Fig.1.12). The small cross-section of the holes reduces further on the measuring volume to much smaller values than permitted by the diffraction limit, in several cases volumes of the order of several attolitres have been reported.

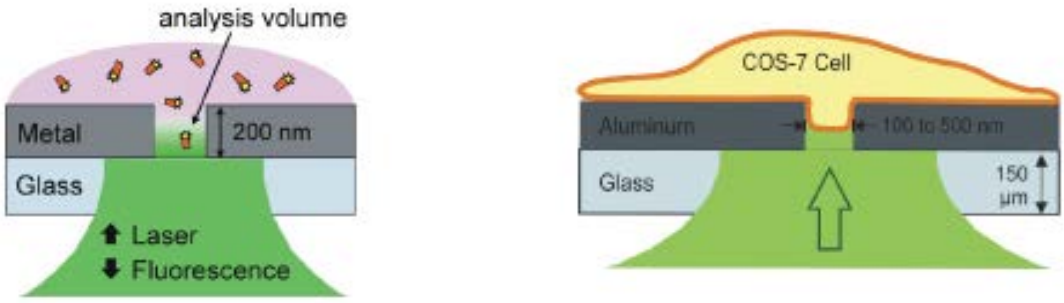
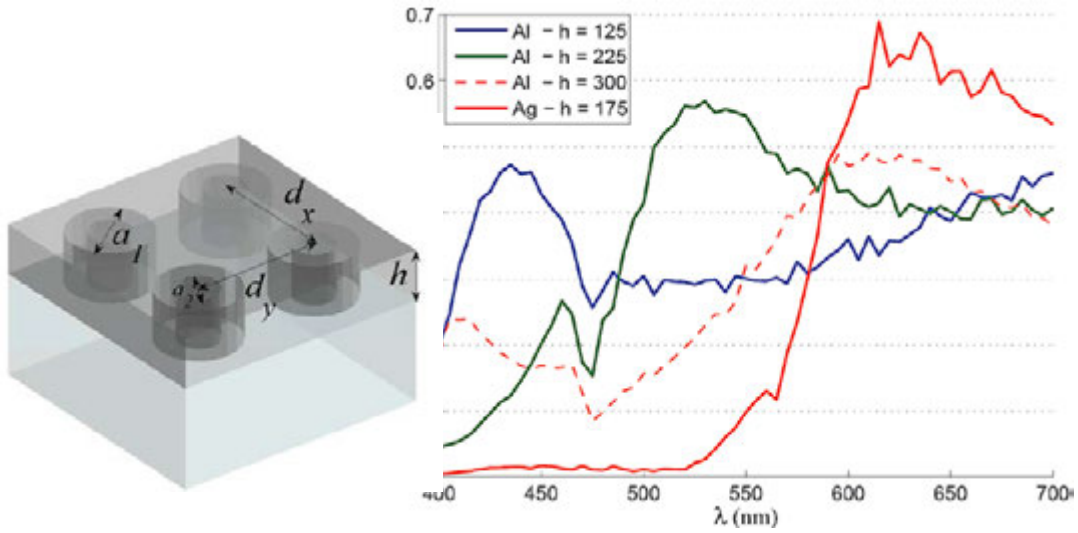


Fig.1.12. Schematic presentation of (a) single-molecule fluorescence backside microscopy field density, and (b) tiny-portion cell membrane microscopy inside a metallic aperture (after [1.28] with the publisher's permission).

1.4.6. Non-resonant filters

Resonant filtering has its advantages, when narrow spectral bands are aimed, but in some important applications it is necessary to have wider bands, accompanied by angular and polarization invariance with respect to the incident wave. Such are the requirements for color filters used to separate the RGB colors on each pixel of the CCD cameras. A promising example [1.29] contains a small subwavelength grating consisting of circular metallic bumps with different diameters and heighst (Fig.1.13a) that can filter light in different spectral regions, depending on their geometrical parameters, (Fig.1.13b).



(a) (b)

Fig.1.13. (a) Nanostucture consisting of coaxial metallic cylinders, with a subwavelength period equal to 300 nm, external diameter 260 nm, and internal diameter 160 nm. (b) Transmission spectrum of the structure for different metals and cylinder heights, as shown in the inset (after [1.29] with the publisher's permission).

1.4.7. Flying natural gratings: butterflies, cicadas

The observation of Hopkinson (see Sec.1.1) of the diffraction on a handkerchief is far the less exotic grating that exists. As always, Nature had all the time to surpass humanity. A striking realization has been developed during the million-year long evolution of butterfly wings that

have so attractive coloring. The so called *Morpho rhetenor* butterflies have a non-pigment metallic blue color (*Fig.1.14*) that long has been attributed to multilayer reflection. However, Vukusic et al. [1.30] had the idea (and funds) to use an electron microscope to observe deeper in detail the structure of the scales, as reported in *Fig.1.15*a, a typical 3D structure that resembles a photonic crystal. The modeling with a 2D grating with a structure given in *Fig.1.15*b confirms the blue-spectrum reflection of the scales (*Fig.1.16*).

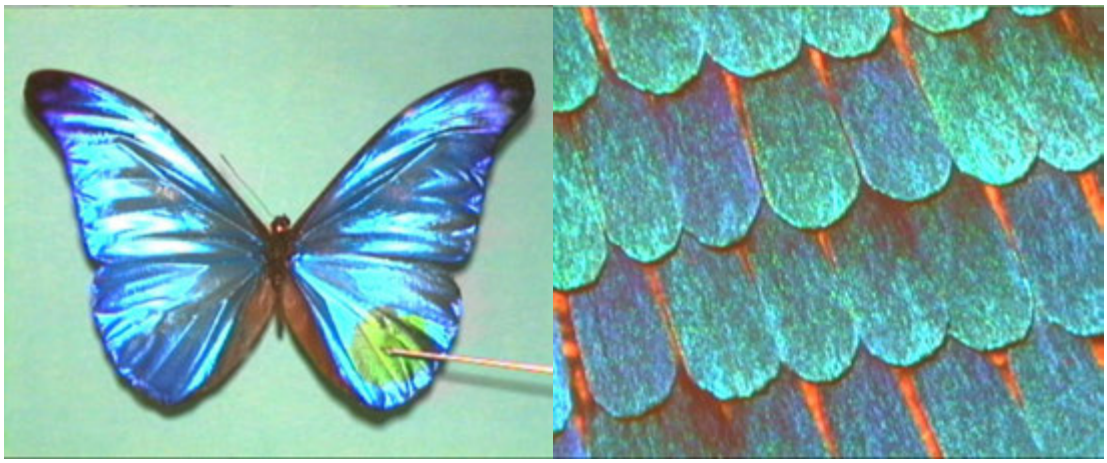


Fig.1.14. Entire view (to the left) and magnification of the scales (to the right) of a Morpho rhetenor butterfly (after [1.30] with the publisher's permission).

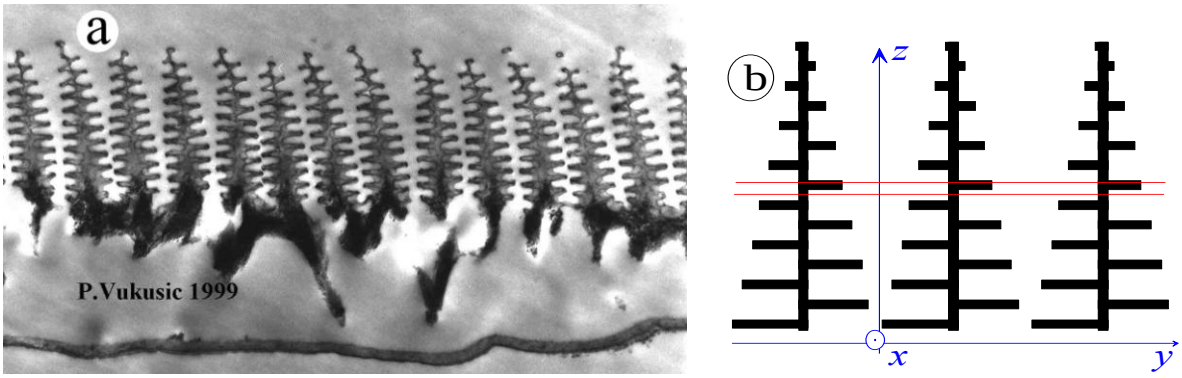


Fig.1.15. (a) Transmission electron microscope image showing the cross-section through a single Morpho rhetenor scale (after [1.30] with the publisher's permission). (b) Modeled structure; the two red lines define a grating layer, the optical index of white regions is 1 (after [1.31] with the publisher's permission).

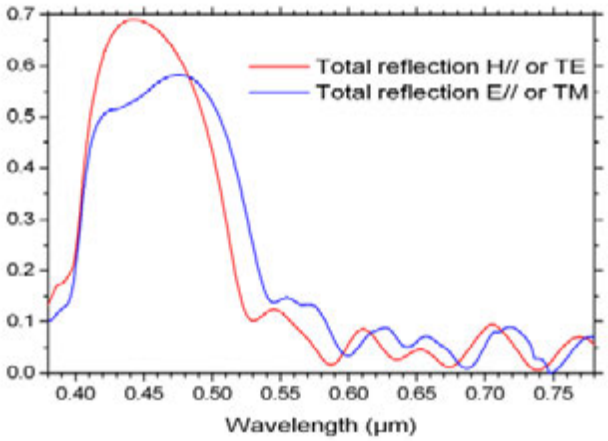
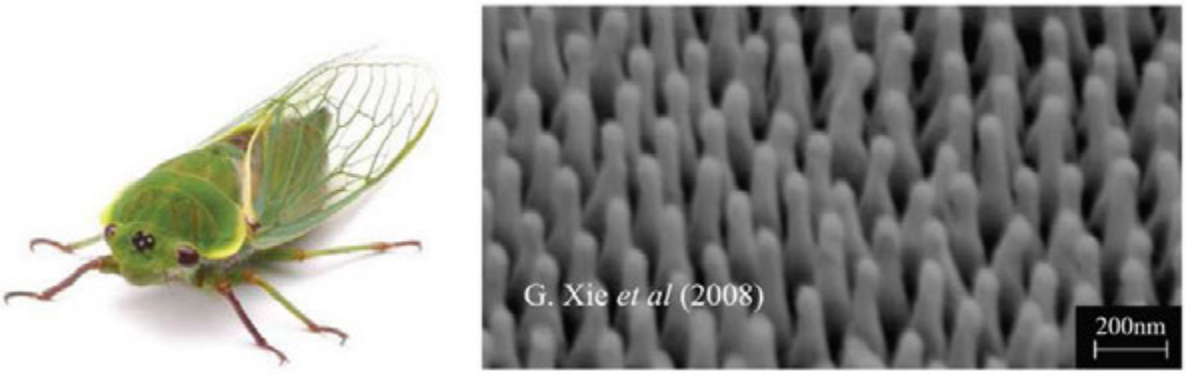


Fig.1.16. Spectral dependence of the reflection of the butterfly scale (after [1.31], with the publisher's permission).

Another "application" is used by the cicadas (quite common in the southern Europe, see Fig.1.17a) to camouflage themselves on the tree branches. Their wings are covered with an anti-reflection nanostructure, first observed by Xie et al. [1.32] and shown in Fig.1.17b. Anyone that has entered the microwave measuring rooms can identify the cones on the walls that are about 1 million times larger, as scaled to the wavelength.



(a) (b)

Fig.1.17. (a) A photo of a cicada, and (b) the nanostructure pattern on its wings (after [1.32] with the publisher's permission).

1.5. Gratings in Integrated optics and plasmonic devices

Gratings are used in integrated optical devices to deviate the direction of propagation of waveguide modes and surface waves, for their focusing inside or outside the guide, or for energy transfer between different modes. Let us consider the case with one-dimensional periodicity. The grating equation can be applied not only to free-space waves, but to the waveguides modes. If the period is suitably chosen, it is possible to couple one mode to another:

$$\mathrm{Re}(k_{g,1}) = \mathrm{Re}(k_{g,2}) - K \tag{1.9}$$

where $k_{g,1}$ and $k_{g,2}$ are the propagation constants of the modes.

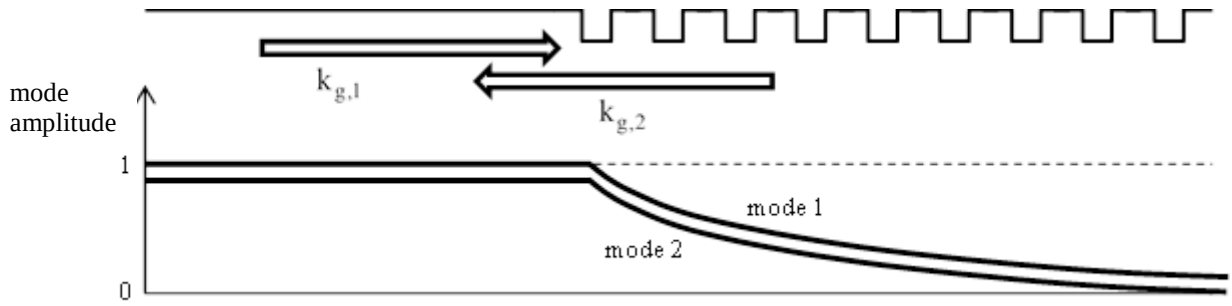


Fig.1.18. Upper part: schematical presentation of Bragg relief type lamellar grating deposited on a waveguide (dielectric or plasmonic) with two propagating modes. Lower part: energy carried by each mode.

It is possible to couple the mode propagating in a given direction to the same but contra-propagative mode. This is the case of the so-called Bragg gratings that act as a distributed mirror forming a forbidden zone for the mode propagation, in which the mode field decreases exponentially without being radiated in the cladding and in the substrate. The result is that it is rejected back into a contra-propagative direction (Fig.1.18). Due to the limit size of the grating region, a small part of the incident mode 1 is transmitted to the right, thus the reflected mode 2 carries smaller amount of energy. The grating grooves can be made curvilinear in order to focus the mode. The same effect can be obtained by replacing the 1D structure by 2D periodicity having also a period in the transversal y-direction (see Section 1.8).

1.6. Beam-splitting applications

The fact that the periodicity creates diffraction orders can be used to create multiple beams from a single laser beam. The most-commonly used device is symmetrical groove transmission gratings used as beam splitters for optical disk readers, where the wavelength of the laser source is constant. As a rule, the zero order beam reads the track and the two first order beams read adjacent tracks to keep the head both centered and focused. By controlling the groove depth, the ratio of zero to first orders transmission can be varied over a factor of 10, and a high degree of symmetry is inherent.

A special type of transmission grating can be used to generate an entire family of orders with a groove shape designed to make their intensities as equal as possible. This gives us multiple beam splitters, which may have 5 or even 20 orders on both sides of zero. A top view of such gratings under working conditions (with a laser input) gives rise to the term of "*fan-out gratings*". Applications are found in scanning reference planes for construction use, optical computing, and others [1.5, 33]. Such gratings tend to have large groove spacings (10 to 100 μm) and low depth modulations. The difficulty in making such gratings lies in achieving a groove shape that leads to a sufficient degree of efficiency uniformity among orders, especially if they are to function over a finite wavelength range. Two obvious candidates are cylindrical sections or an approximation of this shape in the form of a wide angle V with several segments of different angles. The choice may vary with the availability of the corresponding diamond tools. They have also been made by holographic methods, which are able to produce the parabolic groove form that gives the best energy uniformity

between the diffracted orders [1.34, 35]. Applications of fan-out gratings are found in scanning reference planes for construction use, in biophysics for simultaneous treatment of great number of samples, etc.

By combining two such grating at right angles to each other, an accurately defined 2D array of laser beams is generated to be used for calibrating the image field distortion of large precision lenses, in robotic vision systems, or in parallel optical computing. Instead of using two 1D periodical grating, it is possible to use a single large-period 2D crossed grating with specially optimized pattern inside each period [1.36].

1.7. Subwavelength gratings for photovoltaic applications

As explained in Sec.1.4, resonant excitation in metallic gratings can lead to a total absorption of incident light, effect necessary for the efficient work of photovoltaic devices. The problem with surface plasmon excitation and the accompanying light absorption is that its wave is not localized, thus it is characterized by a well-determined value of propagation constant along the surface, because non-local effects in the real space are localized in the inverse space. This is why the surface plasmon anomalies have very narrow angular and spectral width, an advantage in detector construction, but failing in photovoltaics. Evidently, effects that are less localized in the inverse space will be more localized in the real space. This leads us to cavity resonances, volume plasmonic excitation, and surface plasmons that propagate in the vertical direction but are localized in x-y direction. Cavity resonances in deep grooves or in closed cavities, like embedded dielectric spheres or cylinders inside a metallic sheet can absorb light within relatively large angular region (20-30 deg) [1.37-40].

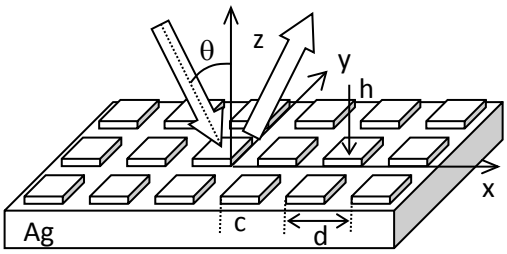


Fig.1.19. Crossed metallic diffraction grating

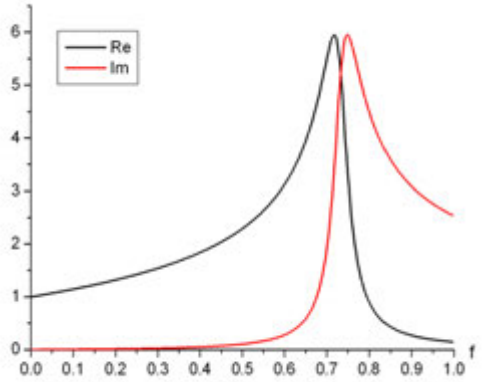


Fig.1.20. Real and imaginary part of the effective refractive index of the structure of Fig.1.19 as a function of the filling ratio when the period is much shorter than the wavelength λ = 457 nm.

However, the problem of cavity resonances is that they are strongly wavelength-sensitive. An alternative approach consists of using metamaterial behavior of small-feature structures. By mixing metallic and dielectric materials one can, in general, obtain strongly absorbing alloys with effective refractive index that does not exist for known materials. In addition, the equivalent metamaterial layer has a uniaxial anisotropy with axis perpendicular to the grating plane. Thus inside the xOy plane it has isotropic properties and its response is polarizationally independent, at least close to normal incidence.

For example, a crossed channel grating as presented in Fig.1.19 and made of silver bumps on a silver substrate. It can strongly absorb the incident light in much larger spectral

domain when compared with the surface plasmon excitation effects. As can be observed in Fig.1.20, close to a filling ratio ($f = c^2/d^2$) equal to 0.7, both the real and the imaginary part of the effective refractive index for small-period structure grow significantly to values that no existing material has in this spectral domain. This increases the effective optical thickness of the system, together with its absorption, so that a very thin grating ($h < 10$ nm) can totally absorbed incident light [1.41]. Because of the 2D periodicity of the structure, it becomes polarization insensible. Moreover, this effect has no resonant nature and is extended angularly to almost the entire set of angles of incidence, as seen in Fig.1.21. In addition, the spectral domain of absorption stronger than 75% extends to a more than 100 nm interval [1.42].

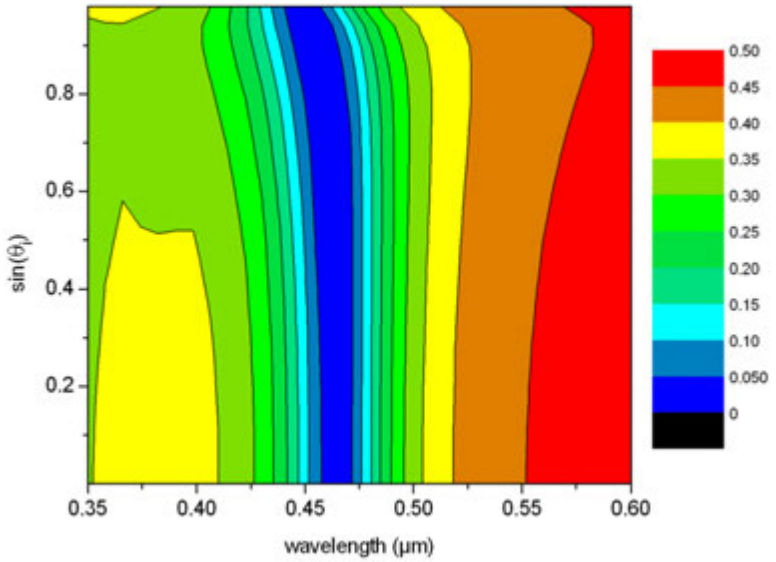


Fig.1.21. Reflectivity of the crossed graing in Fig.1.19 as a function of the wavelength and of the sine of the incident angle. Period d = 3 nm, system thickness h = 6.2 nm, filling factor = 0.7, i.e. c/d = 0.837.

However, the use of such structures for practical applications requires not only a larger spectrum of strong absorption, but also needs that the absorbed light is not immediately transformed into heat and that it is kept inside the photovoltaic layer that transforms light into electricity. The simple structure represented in Fig.1.19 is not sufficient to efficiently fulfill these purposes. It has to be complicated by a structuring in the depth of the device (Fig.1.22). Light is incident from the top of the device. Its entering part (in blue) is designed to act as an antireflector, and absorbs most of the blue part of the spectrum. The green part is absorbed in the middle of the rods, while the back-side part (in red) absorbs the longer wavelengths, and serves as a reflector to prevent the unabsorbed part to be lost in the substrate [1.43].

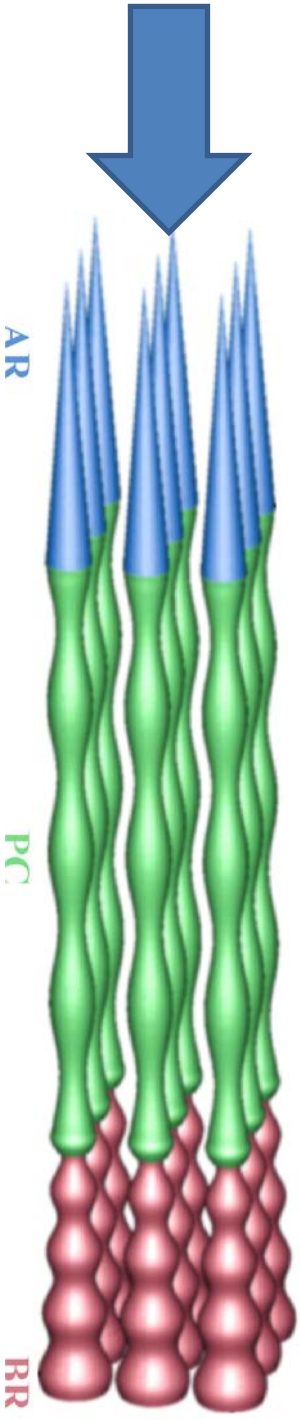


Fig.1.22. Grating rods as optimal photovoltaic absorber (after [1.43], with the publisher's permission)

1.8. Photonic crystals

Periodic structures in optics can serve for the photons in the same manner as semiconductor crystals for electrons. This common feature led in the '90s to call such structures *photonic crystals*. As discovered by Yablonovich [1.44, 45], they present band-gaps that forbid propagation and thus guaranteeing 100% reflection inside the band. While this property is widely known and largely used in multilayer dielectric mirrors (Fig.1.23a), the band gap of 1D photonic crystals is limited in relatively smaller angular interval. Some other structures that have 2D periodicity combined with a nanostructuring in the third dimensions can be found in Figs.1.15, 17, and 22.

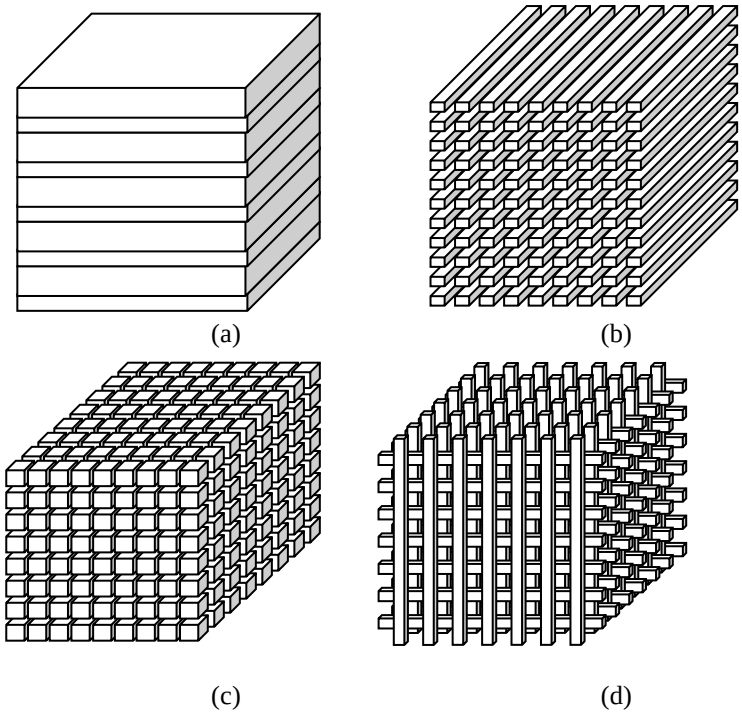


Fig.1.23. Schematic representation of (a) one-, (b) two- and (c, d) three-dimensional photonic crystals.

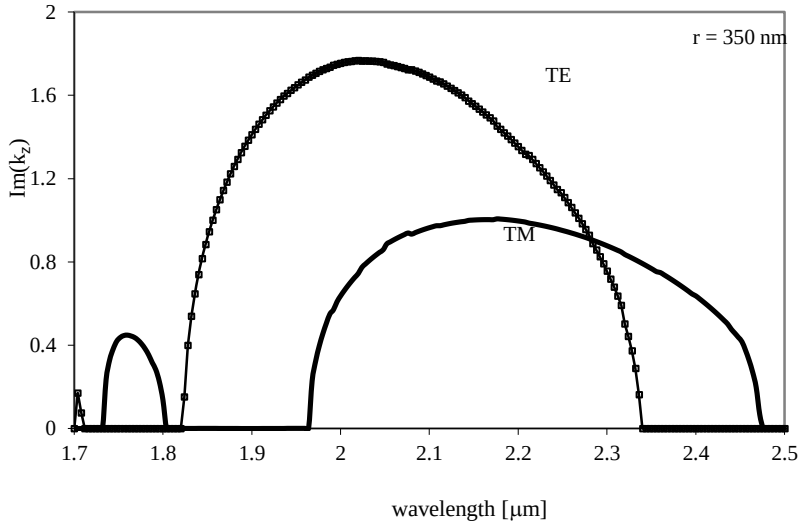


Fig.1.24. Forbidden bands for the photonic crystalmade of circular cylinders with d = 1.414 μm, r = 0.35 μm (after [1.46] with the publisher's permission).

A typical band-gap structure of a 2D photonic crystal presented in Fig.1.23b, but its surface cut at an angle of 45° with respect to the vertical direction, is given in Fig.1.24 for a system having period d = 1.414 μm in both directions, and consists of circular cylindrical rods with radius equal to 0.35 μm and optical index of 2.9833 in air as a matrix. The figure presents the values of the smallest imaginary part of k_z. As can be seen, a forbidden gap in both TE and TM polarization exists for $\lambda \in [1.97, 2.33\,\mu m]$ [1.46, 47]. Inside the band gap electromagnetic field intensity diminishes exponentially and the entire incident light is reflected back. However, the choice of the ratio of the wavelength and the period allows for the propagation of the -1st order in reflection. This gives the possibility to guide the entire incident light into this order, thus perfect blazing in the -1st diffracted order can be obtained in both polarizations, a property that is strongly desirable in many applications. And indeed, Fig.1.25 presents the diffraction efficiency of the system having cylinders with radii of 350 nm (a) and 150 nm (b). A well-defined spectral region with almost 100% efficiency in unpolarized light can be observed.

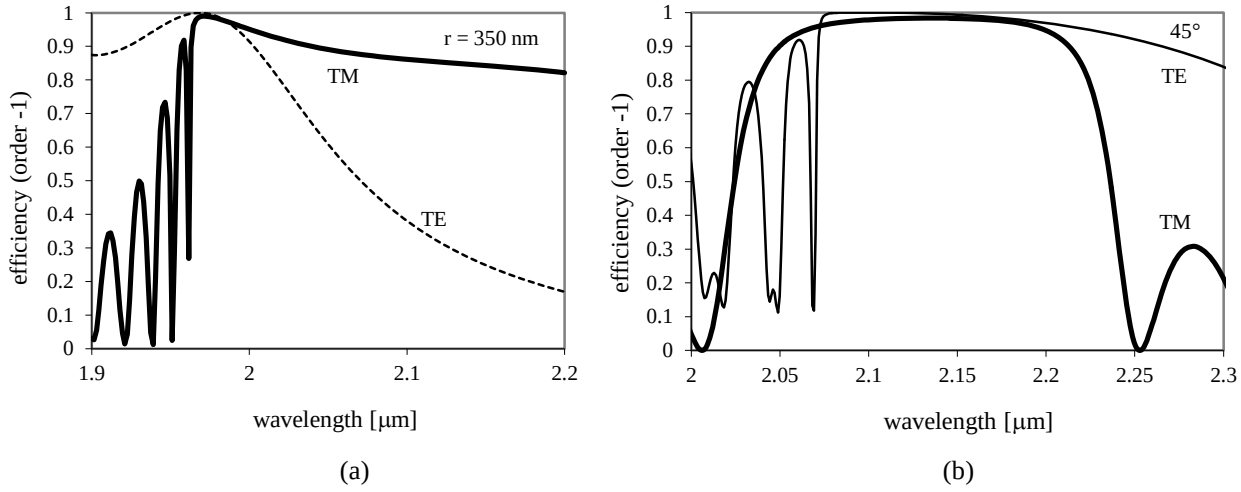


Fig.1.25. Diffraction efficiency in order -1 as a function of the wavelength lying inside the band gap. (a) r = 350 nm, (b) r = 150 nm (after [1.46] with the publisher's permission).

The property of totally reflecting the incident light whatever the direction, can be of great importance for light guiding and manipulation. Light confinement and guiding in a single dimension is ensured by using planar waveguides. Channel waveguides and optical fibers confine light in two dimensions, but they suffer from two important limitations: dispersion and bending losses. High bending angles damage the guiding properties and lead to radiation losses. A waveguide constructed with photonic crystal walls can ensure bending, Fig.1.26, without losses even at 90°, as predicted numerically [1.48].

It is impossible even only to list here the *potential* applications of photonic crystal devices, as for example negative refraction, perfect lenses construction, photonic crystal fibers, nonlinear optical applications. The main problem that persists is purely *technological*: while it is relatively easy to fabricate 3D periodically structures working in the microwave and far-IR domain, scaling down to the visible and the near-IR presents a lot of challenges to optical industry. Fortunately, the process of fiber manufacturing enables literally such mechanical scaling of the dimension, transferring the initial large-diameter preform into a thin fiber by pulling it. If the preform is carefully drilled with macroscopic holes, they are preserved in the final fibers, but scaled to nanodimensions.

In the resulting photonic crystal fiber (Fig.1.27), light is preserved in the central hollow guide by repulsion from the surrounding structure that presents a forbidden gap for the working wavelength.

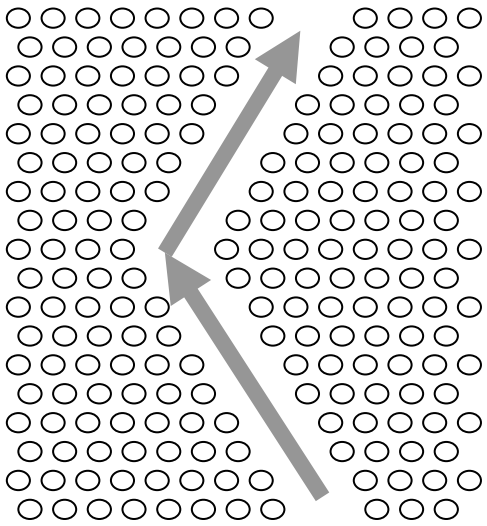


Fig.1.26. Light guiding in a cavity formed by photonic crystal walls consisting of cylindrical objects

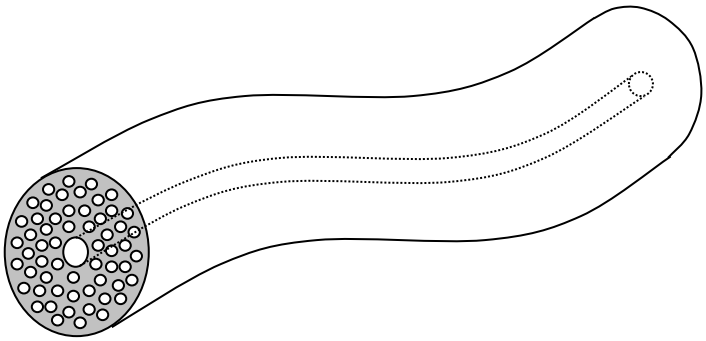


Fig.1.27. Schematic representation of a portion of an optical fiber – photonic crystal hybrid structure. Small cylindrical holes run along the fiber length. A central hole (shown with a dashed line inside the fiber) serves as an energy propagator, which ensured low dispersion, low absorption losses and high damage threshold

1.9. References:

1. *The Concise Oxford Dictionary of English Etymology in English Language Reference* accessed via Oxford Reference Online
2. D. Rittenhouse "An optical problem proposed by F. Hopkinson and solved," J. Am. Phil. Soc. **201**, 202-206 (1786)
3. T. Young: "On the theory of light and colors," Phil. Trans. **II**, 399-408 (1803)
4. J. Fraunhofer: "Kurtzer Bericht von the Resultaten neuerer Versuche über die Gesetze des Lichtes, und die Theorie derselbem," Gilberts Ann. Phys. **74**, 337-378 (1823); "Über die Brechbarkeit des electrishen Lichts," K. Acad. d. Wiss. zu München, April-June 1824, pp.61-62
5. E. Loewen and E. Popov, : *Diffraction Gratings and Applications*, (Marcel Dekker, New York, 1997)
6. J. Hoose and E. Popov, "Two-dimensional gratings for low polarization dependent wavelength demultimlexing," Appl. Opt. **47**, 4574 – 4578 (2008)
7. E. Popov, J. Hoose, B. Frankel, C. Keast, M. Fritze, T.Y. Fan, D. Yost, and S. Rabe: "Diffraction-grating based wavelength demultimlexer," Opt. Expr. 12, 269-275 (2004)
8. J. Hoose, R. Frankel, E. Popov, and M. Nevière: "Grating device with high diffraction efficiency," US patent No 6958859/25.10.2005
9. E. Popov, B. Bozhkov, D. Maystre, and J. Hoose: "Integral method for echelles covered with lossless or absorbing thin dielectric layers," Appl. Opt. **38**, 47-55 (1999)
10. H. U. Käufl, "N-band long slit grism spectroscopy with TIMMI at the 3.6 m telescope," The ESO Messenger, **78**, 4 -7 (Dec. 1994)
11. K. Knop, "Diffraction gratings for color filtering in the zero diffracted order," Appl. Opt. **17**, 3598 – 3603 (1978)
12. Lord Rayleigh, "Note on the Remarkable Case of Diffraction Spectra Described by Prof. Wood," Philos. Mag. **14**, 60 (1907)
13. R. W. Wood, "On a remarkable case of uneven distribution of light in a diffraction grating spectrum," Phylos. Mag. **4**, 396-402 (1902)
14. U. Fano, "The theory of anomalous diffraction gratings and of quasi-stationary waves on metallic surfaces (Sommerfeld's waves)," J. Opt. Soc. Am. **31**, 213-222 (1941)
15. A. Hessel and A. A. Oliner, "A new theory of Wood's anomalies on optical gratings," Appl. Opt. **4**, 1275-1297 (1965)
16. M. C. Hutley and D. Maystre, "Total absorption of light by a diffraction grating," Opt. Commun. **19**, 431-436 (1976)
17. M. J. Jory, P. S. Vukusic, and J. R. Sambles, "Development of a prototype gas sensor using surface plasmon resonance on gratings," Sens. Actuators B **17,** 203-209 (1994)
18. N. Bonod, E. Popov, and R. C. McPhedran, "Increased surface plasmon resonance sensitivity with the use of double Fourier harmonic gratings," Opt. Express **16**, 11691-11702 (2008)
19. L. Mashev and E. Popov: "Zero Order Anomaly of Dielectric Coated Grating," Opt. Commun. **55**, 377 (1985)
20. G. A. Golubenko, A. S. Svakhin, V. A. Sychugov, A. V. Tischenko, E. Popov and L. Mashev: "Diffraction Characteristics of Planar Corrugated Waveguides," J. Opt. Quant. Electr. **18**, 123 (1986)
21. E. Popov, L. Mashev and D. Maystre: "Theoretical Study of the Anomalies of Coated Dielectric Gratings," Opt. Acta **33**, 607 (1986)
22. S. Wang, R. Magnusson, J. Bagdy, and M. Moharam, "Guided-mode resonances in planar dielectric-layer diffraction gratings," J. Opt. Soc. A **7**, 1470-1474 (1990)
23. S. Tibuleac and R. Magnusson, "Narrow-linewidth bandpass filters with diffractive thin-film layers," Opt. Lett. **26**, 584-586 (2001)

24. A. Sentenac and A.-L. Fehrembach, “Angular tolerant resonant grating filters under oblique incidence,” J. Opt. Soc. Am. A, **22**, 475-480 (2005)
25. T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, P. A. Wolff, “Extraordinary optical transmission through subwavelength hole arrays”, Nature, **391**, 667-669 (1998)
26. E. Popov, M. Neviere, S. Enoch, and R. Reinisch, “Theory of light transmission through subwavelength periodic hole arrays,” Phys. Rev. B, **62**, 16100-16108 (2000)
27. S. Enoch, E.Popov, M.Nevière, and R. Reinisch, “Enhanced light transmission by hole arrays,” J. Opt. A: Pure Appl. Opt. **4**, S83-S87 (2002)
28. J. Wenger, D. Gerard, P.-F. Lenne, H. Rigneault, N. Bonod, E. Popov, D. Marguet, C. Nelep, T. Ebbesen, “Biophotonics applications of nanometric apertures,” Int. J. Materials and Product Technology. **34**, 488-506 (2009)
29. G. Demésy, “Modélisation électromagnétique tri-dimensionnelle de réseaux complexes. Application au filtrage spectral dans les imageurs CMOS”, Ph. D. Thèse 2009AIX300006, Univ. Aix-Marseille III, Marseille (2009)
30. P. Vukusic, J.R. Sambles, C.R. Lawrence and R.J. Wootton, "Quantified interference and diffraction in single *Morpho* butterfly scales,"Proceedings: Biological Sciences, The Royal Society of London **266**, 1403 -1411 (1999)
31. B. Gralak, G. Tayeb, and S. Enoch, “Morpho butterflies wings color modeled with lamellar grating theory,” Opt. Express **9**, 567-578 (2001)
32. G. Xie, G. Zhang, F. Lin, J. Zhang, Z. Liu, and S. Mu, “The fabrication of subwavelength anti-reflective nanostructures using a bio-template,” *IOP Nano*, **19**, 95605, (2008)
33. L. P. Boivin: "Multiple imaging using various types of simple phase gratings," Appl. Opt., **11**, 1782–1792 (1972).
34. P. Langlois and R. Beaulieu, “Phase relief gratings with conic section profile in the production of multiple beams,” Appl. Opt. **29**, 3434-3439 (1990).
35. D. Shin and R. Magnusson: "Diffraction of surface relief gratings with conic cross–sectional gratings shapes," J. Opt. Soc. Am. A **6**, 1249–1253 (1989).
36. NOI Bulletin, v.**5**, no.2, July 1994, Québec, Canada
37. N. Bonod and E. Popov, “Total light absorption in a wide range of incidence by nanostructured metals without plasmons,” Opt. Lett. **33**, 2287-2289 (2008)
38. E. Popov, L. Tsonev, and D. Maystre, ‘Lamellar metallic grating anomalies,” Appl. Opt. **33**, 5214-5219 (1994)
39. E. Popov, N. Bonod, and S. Enoch, “Comparison of plasmon surface wave on shallow and deep 1D and 2D gratings,” Opt. Express **15**, 4224-4237 (2007)
40. E. Popov, D. Maystre, R. C. McPhedran, M. Nevière, M. C. Hutley, and G. H. Derrick, "Total absorption of unpolarized light by crossed gratings," Opt. Express **16**, 6146-6155 (2008)
41. J. Le Perchec, P. Quémerais, A. Barbara, and T. López-Rios, “Why metallic surfaces with grooves a few nanometers deep and wide may strongly absorb visible light,” Phys. Rev. Lett. **100**, 066408 (2008)
42. E. Popov, S Enoch, and N. Bonod, “Absorption of light by extremely shallow metallic gratings: metamaterials behavior,” Opt. Express **17**, 6770-6781 (2009)
43. G. Demésy and S. John, “Solar energy trapping with modulated silicon nanowire photonic crystal,” J. Appl. Phys., 074326 (2012)
44. E. Yablonovitch, “Inhibited spontaneous emission in solid-state physics and electronics,” Phys. Rev. Lett. **58**, 2059-2062 (1987)
45. E. Yablonovitch, “Photonic crystals,” J. Modern Opt., **41**,173-194 (1994)
46. E. Popov, B. Bozhkov, and M. Nevière, “Almost perfect blazing by photonic crystal rod gratings,” Appl. Opt. **40**, 2417-2422 (2001)

47. E. Popov and B. Bozhkov: “Differential method applied for photonic crystals”, Appl. Opt. **39**, 4926 (2000)
48. G. Tayeb and D. Maystre, “Rigorous theoretical study of finite size two-dimensional photonic crystals doped by microcavities,” J. Opt. Soc. Am. A **14**, 3323-3332 (1997)

Chapter 2:
Analytic Properties of Diffraction Gratings

Daniel Maystre

Table of Contents:

Chapter 2

Analytic Properties of Diffraction Gratings

Daniel Maystre

CNRS, Aix Marseille Université, Centrale Marseille, Institut Fresnel UMR 7249,
Campus Universitaire de Saint Jérôme
13397 Marseille Cedex 20, France
daniel.maystre@fresnel.fr

2.1 Introduction

Since the 80's, specialists of gratings can rely on very powerful grating softwares [1-6]. These softwares are able to compute grating efficiencies for almost any kind of grating in any domain of wavelength, even though the progress of grating technologies needs endless extensions of grating theories to new kinds of structures. These softwares are based on elementary laws of Electromagnetics. Using mathematics, these laws lead to boundary value problems which can be solved on computers using adequate algorithms.

However, a grating user should not ignore some general properties of gratings which can derived directly from the boundary value problem without any use of computer. These analytic properties are valuable at least for two reasons. First, they strongly contribute to a better understanding of an instrument which puzzled and fascinated many specialists of Optics since the beginning of the 20th century. Secondly, they allow a theoretician to check the validity of a new theory or its numerical implementation, although one must be very cautious: a theory can fail while its results satisfy some analytic rules. Specially, this surprising remark apply to properties like energy balance or reciprocity theorem.

The first part of this chapter is devoted to the use of the elementary laws of Electromagnetics for stating the boundary value problems of gratings in various cases of materials and polarizations. Then, we deduce from the boundary value problems the most important analytic properties of gratings.

2.2 From the laws of Electromagnetics to the boundary-value problems

2.2.1 Presentation of the grating problem

Figure 2.1 represents a diffraction grating. Its periodic profile $\mathscr{P}$ of period d along the x axis separates air (region $\mathscr{R}_0$) from a grating material (region $\mathscr{R}_1$) which is generally a metal or a dielectric. The y axis is the axis of invariance of the structure and the z axis is perpendicular to the average profile plane. We denote by z_M the ordinate of the top of $\mathscr{P}$, its bottom being located

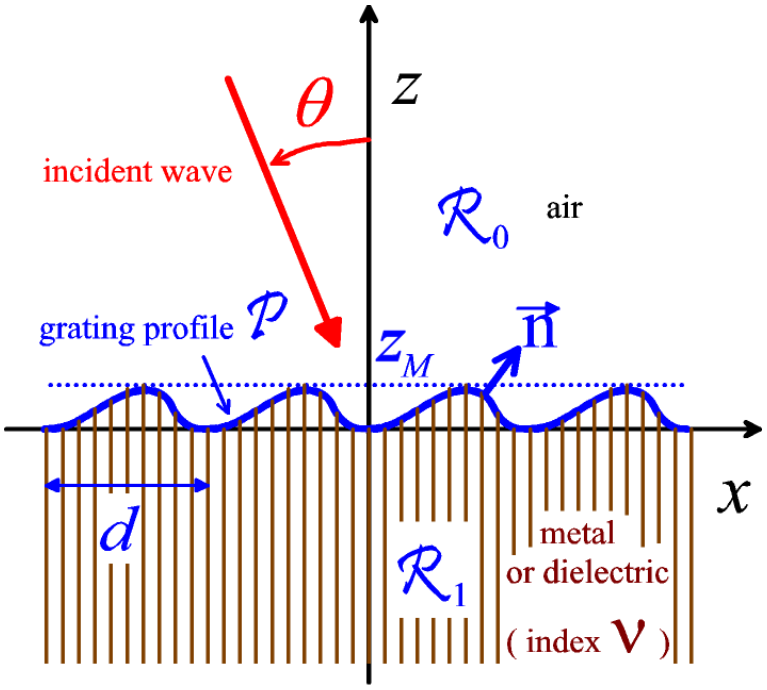


Figure 2.1: Notations.

on the xy plane by hypothesis. We suppose that the incident light can be described by a sum of monochromatic radiations of different frequencies. Each of these can in turn be described in a time-harmonic regime, which allows us to use the complex notation (with an $\exp(-i\omega t)$ time-dependence). In this chapter, we assume that the wave-vector of each monochromatic radiation lies in the cross-section of the grating (xz plane). In the following, we deal with a single monochromatic radiation.

The electromagnetic properties of the grating material (assumed to be non-magnetic) are represented by its complex refractive index ν which depends on the wavelength $\lambda = 2\pi c/\omega$ in vacuum ($c = 1/\sqrt{\varepsilon_0\mu_0}$ being the speed of light, with ε_0 and μ_0 the permittivity and the permeability of vacuum). This complex index respectively includes the conductivity (for metals) and/or the losses (for lossy dielectrics). It becomes a real number for lossless dielectrics.

In the air region, the grating is illuminated by an incident plane wave. The incident electric field $\overrightarrow{E^i}$ is given by :

$$\overrightarrow{E^i} = \overrightarrow{P}\exp\big(ik_0 x\sin(\theta) - ik_0 z\cos(\theta)\big), \tag{2.1}$$

with θ being the angle of incidence, from the z axis to the incident direction, measured in the counterclockwise sense, and k_0 being the wavenumber in the air ($k_0 = 2\pi/\lambda$, we take an index equal to unity for air). The wave-vector of the incident wave is given by:

$$\overrightarrow{k_0^i} = \begin{bmatrix} k_0\sin(\theta) \\ 0 \\ -k_0\cos(\theta) \end{bmatrix}. \tag{2.2}$$

The physical problem is to find the total electric and magnetic fields $\overrightarrow{E}$ and $\overrightarrow{H}$ at any point of space.

2.2.2 Maxwell's equations

First, let us notice that the physical problem remains unchanged after translations of the grating or of the incident wave along the y axis since they do not depend on y. Therefore, if $\overrightarrow{E}(x,y,z)$ and $\overrightarrow{H}(x,y,z)$ are the total fields for a given grating and a given incident wave, $\overrightarrow{E}(x,y+y_0,z)$ and $\overrightarrow{H}(x,y+y_0,z)$ will be solutions too, regardless of the value of y_0. Assuming, from the physical intuition, that the solution of the grating problem is unique, we deduce that $\overrightarrow{E}$ and $\overrightarrow{H}$ are independent of y.

In order to state the mathematical problem, we use the harmonic Maxwell equations in $\mathscr{R}_0$:

$$\nabla \times \overrightarrow{E} = i\omega\mu_0 \overrightarrow{H}, \tag{2.3}$$

$$\nabla \times \overrightarrow{H} = -i\omega\varepsilon \overrightarrow{E}, \tag{2.4}$$

with:

$$\varepsilon = \begin{cases} \varepsilon_0 & \text{in} \quad \mathscr{R}_0, \\ \varepsilon_1 = \varepsilon_0 \nu^2 & \text{in} \quad \mathscr{R}_1. \end{cases} \tag{2.5}$$

In the following, equations (2.3) and (2.4) will be called first and second Maxwell equations respectively. We note that Maxwell's equations $\nabla.\overrightarrow{E} = 0$ and $\nabla.\overrightarrow{H} = 0$ are the straightforward consequences of the first and second Maxwell equations (it suffices to take the divergence of both members).

We introduce the diffracted fields $\overrightarrow{E^d}$ and $\overrightarrow{H^d}$ defined by:

$$\overrightarrow{E^d} = \begin{cases} \overrightarrow{E} - \overrightarrow{E^i} & \text{in} \quad \mathscr{R}_0, \\ \overrightarrow{E} & \text{in} \quad \mathscr{R}_1, \end{cases} \tag{2.6}$$

$$\overrightarrow{H^d} = \begin{cases} \overrightarrow{H} - \overrightarrow{H^i} & \text{in} \quad \mathscr{R}_0, \\ \overrightarrow{H} & \text{in} \quad \mathscr{R}_1. \end{cases} \tag{2.7}$$

The interest of the notion of diffracted field is that it satisfies the so-called **radiation condition** (or Sommerfeld condition, or outgoing wave condition), in contrast with the total field which does not satisfy this condition in $\mathscr{R}_0$ since it includes the incident field. This means that the diffracted fields must remain bounded and propagate upwards in $\mathscr{R}_0$ when $z \to +\infty$. The same property must be satisfied in $\mathscr{R}_1$, but that time the diffracted fields must remain bounded and propagate downwards in $\mathscr{R}_1$ when $z \to -\infty$. Since the incident fields satisfy Maxwell's equations in $\mathscr{R}_0$, the diffracted fields satisfy these equations as well. Introducing the components of the diffracted fields on the three axes, Maxwell's equations yield:

$$\partial E_y^d/\partial z = -i\omega\mu_0 H_x^d, \tag{2.8a}$$

$$\partial E_y^d/\partial x = i\omega\mu_0 H_z^d, \tag{2.8b}$$

$$\partial E_z^d/\partial x - \partial E_x^d/\partial z = -i\omega\mu_0 H_y^d, \tag{2.8c}$$

$$\partial H_y^d/\partial z = i\omega\varepsilon E_x^d, \tag{2.9a}$$

$$\partial H_y^d/\partial x = -i\omega\varepsilon E_z^d, \tag{2.9b}$$

$$\partial H_z^d/\partial x - \partial H_x^d/\partial z = i\omega\varepsilon E_y^d. \tag{2.9c}$$

2.2.3 Boundary conditions on the grating profile

On the grating profile, the tangential component of the electric and magnetic fields must be continuous[1]. Thus the boundary condition is given by:

$$(\overrightarrow{[E^d]_0} + \overrightarrow{[E^i]_0}) \times \overrightarrow{n} = \overrightarrow{[E^d]_1} \times \overrightarrow{n}, \tag{2.10}$$

$$(\overrightarrow{[H^d]_0} + \overrightarrow{[H^i]_0}) \times \overrightarrow{n} = \overrightarrow{[H^d]_1} \times \overrightarrow{n}, \tag{2.11}$$

with $\overrightarrow{n}$ being the unit normal to $\mathscr{P}$, oriented toward region $\mathscr{R}_0$ (figure 2.1) and the symbol $[\overrightarrow{F}]_p$ denoting the limit of $\overrightarrow{F}$ when a point of region $\mathscr{R}_p$ tends to the grating profile (with $p \in (0,1)$). As for Maxwell's equations, we note that the other boundary conditions on the normal components of the fields are consequences of equations (2.10) and (2.11). It is worth noting that the linkage between these two boundary conditions is a typical example of an elementary property which is difficult to establish, at least for those who are not acquainted with the theory of distributions. Projecting equations (2.10) and (2.11) on the three axes yields:

$$[E_y^d]_0 - [E_y^d]_1 = -[E_y^i]_0, \tag{2.12a}$$

$$n_x[E_z^d]_0 - n_z[E_x^d]_0 - n_x[E_z^d]_1 + n_z[E_x^d]_1 = -n_x[E_z^i]_0 + n_z[E_x^i]_0, \tag{2.12b}$$

$$[H_y^d]_0 - [H_y^d]_1 = -[H_y^i]_0, \tag{2.13a}$$

$$n_x[H_z^d]_0 - n_z[H_x^d]_0 - n_x[H_z^d]_1 + n_z[H_x^d]_1 = -n_x[H_z^i]_0 + n_z[H_x^i]_0. \tag{2.13b}$$

2.2.4 Separating the general boundary-value problem into two separated scalar problems

The first conclusion to draw from equations (2.8), (2.9), (2.12) and (2.13) is that **they can be separated into two independent sets**. The first one, called TE case, includes equations (2.8a), (2.8b), (2.9c), (2.12a) and (2.13b). It only contains the transverse component (viz. the y-component) E_y^d of the electric field and the xz components (orthogonal to the y axis) H_x^d and H_z^d of the magnetic field. It must be remembered that the incident field $\overrightarrow{E^i}$ is given by equation (2.1) and thus is not an unknown field. The same remark applies to the complementary set (TM case), but with the transverse component of the magnetic field and the xz components of the electric field. As a consequence, the general problem of diffraction by a grating can be decomposed into two elementary mathematical problems.

2.2.4.1 The TE case problem

In the first one, the xz components of the magnetic field can be expressed as functions of the transverse component of the electric field using equations (2.8a) and (2.8b). Inserting their expression in equation(2.9c) shows that E_y^d satisfies a Helmholtz equation:

$$\boxed{\nabla^2 E_y^d + k^2 E_y^d = 0,} \tag{2.14}$$

[1]The continuity of the tangential component of the magnetic field is valid for materials having bounded values of permittivity. When the permittivity of the grating material is infinite, as in the model of perfectly conducting material, this condition does not hold.

with:

$$k = \begin{cases} k_0 & \text{in} \quad \mathscr{R}_0, \\ k_1 = k_0 \nu & \text{in} \quad \mathscr{R}_1. \end{cases} \tag{2.15}$$

The associated boundary condition on the diffracted electric field can be deduced from equations (2.12a) and (2.1):

$$\boxed{\left[E_y^d\right]_0 - \left[E_y^d\right]_1 = -P_y \exp\big(ik_0 x \sin(\theta) - ik_0 z \cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P},} \tag{2.16}$$

while the associated boundary condition on its normal derivative can be deduced from equations (2.13b), (2.8a) and (2.8b):

$$\boxed{\begin{aligned} &\left[\frac{\mathrm{d}E_y^d}{\mathrm{d}n}\right]_0 - \left[\frac{\mathrm{d}E_y^d}{\mathrm{d}n}\right]_1 = -\left[\frac{\mathrm{d}E_y^i}{\mathrm{d}n}\right]_0, \\ &= -iP_y \overrightarrow{n}.\overrightarrow{k_0^i} \exp\big(ik_0 x \sin(\theta) - ik_0 z \cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P}, \end{aligned}} \tag{2.17}$$

with $\dfrac{\mathrm{d}F}{\mathrm{d}n}$ denoting the normal derivative $\overrightarrow{n}.\nabla F$. It can be noticed that equation (2.17) entails the continuity of the normal derivative of the transverse component of the total electric field. Equations (2.14), (2.16) and (2.17) are not sufficient to define the boundary-value problem for TE case. A fourth condition must be added: the radiation condition:

$$\boxed{E_y^d \text{ must satisfy a radiation condition for } z \to \pm\infty.} \tag{2.18}$$

The boundary value problem allows us to deduce a fundamental property of gratings. Let us suppose that the incident field is TE polarized, i.e. that the electric incident field is parallel to the y axis ($P_x = P_z = 0$). In these conditions, the equations associated with the TM case are homogeneous: they do not contain the incident field since the right-hand member of equation (2.12b) vanishes. If we believe that the solution of the grating problem is unique, it must be concluded that the xz component of the diffracted and total electric field vanish. On the other hand, the magnetic field is parallel to the xz plane. In other words, **in the TE case, the grating problem becomes scalar: we must determine the y-component of the diffracted electric field**. The xz components of the magnetic field deduce the y-component of the diffracted electric field using equations (2.8a) and (2.8b).

2.2.4.2 The TM case problem

Now, let us deal with the TM case. As for the TE case, it can be shown that the y-component of the magnetic field satisfies a Helmholtz equation by using equations (2.8c), (2.9a) and (2.9b):

$$\boxed{\nabla^2 H_y^d + k^2 H_y^d = 0.} \tag{2.19}$$

The boundary conditions need the calculation of the incident magnetic field. From equation (2.1) and Maxwell equation (2.3), it turns out that:

$$\overrightarrow{H^i} = \overrightarrow{Q} \exp\big(ik_0 x \sin(\theta) - ik_0 z \cos(\theta)\big), \tag{2.20}$$

with:

$$\overrightarrow{Q} = \frac{1}{\omega\mu_0}\overrightarrow{k_0^i}.\overrightarrow{P}\exp\big(ik_0x\sin(\theta) - ik_0z\cos(\theta)\big). \tag{2.21}$$

The associated boundary condition on the diffracted magnetic field can be deduced from equations (2.13a) and (2.20):

$$\boxed{[H_y^d]_0 - [H_y^d]_1 = -Q_y\exp\big(ik_0x\sin(\theta) - ik_0z\cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P},} \tag{2.22}$$

while the boundary condition on its normal derivative is obtained by inserting the expressions of the xz components of the electric field (equations (2.9a) and (2.9b)) in equation (2.12b). Remarking that the incident field satisfies the same equations, we obtain finally:

$$\boxed{\begin{aligned}&\frac{1}{\varepsilon_0}\left[\frac{\mathrm{d}H_y^d}{\mathrm{d}n}\right]_0 - \frac{1}{\varepsilon_1}\left[\frac{\mathrm{d}H_y^d}{\mathrm{d}n}\right]_1 = -\frac{1}{\varepsilon_0}\left[\frac{\mathrm{d}H_y^i}{\mathrm{d}n}\right]_0,\\ &= -\frac{iQ_y}{\varepsilon_0}\overrightarrow{n}.\overrightarrow{k_0^i}\exp\big(ik_0x\sin(\theta) - ik_0z\cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P}.\end{aligned}} \tag{2.23}$$

It can be noticed that equation (2.23) has a simple interpretation: the product $\dfrac{1}{\varepsilon}\dfrac{\mathrm{d}H_y}{\mathrm{d}n}$ is continuous across the profile. Finally, the radiation condition yields:

$$\boxed{H_y^d \text{ must satisfy a radiation condition for } z \to \pm\infty.} \tag{2.24}$$

Equations (2.19), (2.22), (2.23) and radiation conditions for $z \to \pm\infty$ define the boundary-value problem for TM case. As for TE case, the uniqueness of the solution shows that that when the magnetic incident field is parallel to the y axis ($Q_x = Q_z = 0$, the equations associated with the TE case are homogeneous: they do not contain the incident field. It can be concluded that the xz components of the diffracted and total magnetic fields vanish. On the other hand, the electric field is parallel to the xz plane. In other words, **in the TM case, the grating problem becomes scalar: we must determine the y-component of the diffracted magnetic field**. The xz components of the electric field deduce from the y-component of the diffracted magnetic field using equations (2.9a) and (2.9b).

2.2.4.3 TE and TM cases: a unified presentation of the boundary-value problem

In order to deal with both cases simultaneously, we denote by F^d the field defined by:

$$F^d = \begin{cases} E_y^d & \text{for TE case,} \\ H_y^d & \text{for TM case.} \end{cases} \tag{2.25}$$

In the same way, by assuming that the incident field has a unit amplitude (P_y=1 for TE case and Q_y=1 for TM case), the incident field in both cases is given by:

$$F^i = \exp\big(ik_0x\sin(\theta) - ik_0z\cos(\theta)\big), \tag{2.26}$$

the total field F being given by:

$$F = \begin{cases} F^d + F^i & \text{in} \quad \mathscr{R}_0, \\ F^d & \text{in} \quad \mathscr{R}_1. \end{cases} \tag{2.27}$$

Using equations (2.14), (2.16), (2.17), (2.18), (2.19), (2.22), (2.23) and (2.24), it is possible to gather both cases in a unique set of equations:

$$\nabla^2 F^d + k_0^2 F^d = 0, \tag{2.28}$$

$$\left[F^d\right]_0 - \left[F^d\right]_1 = -\exp\left(ik_0 x\sin(\theta) - ik_0 z\cos(\theta)\right) \quad \text{with } (x,z) \in \mathscr{P}, \tag{2.29}$$

$$\frac{1}{\tau_0}\left[\frac{\mathrm{d}F^d}{\mathrm{d}n}\right]_0 - \frac{1}{\tau_1}\left[\frac{\mathrm{d}F^d}{\mathrm{d}n}\right]_1, \quad = -\frac{i}{\tau_0}\overrightarrow{n}.\overrightarrow{k_0^i}\exp\left(ik_0 x\sin(\theta) - ik_0 z\cos(\theta)\right), \quad \text{with } (x,z) \in \mathscr{P}, \tag{2.30}$$

$$F^d \text{ must satisfy a radiation condition for } y \to \pm\infty, \tag{2.31}$$

with:

$$\tau_i = \begin{cases} 1 & \text{for TE case,} \\ \varepsilon_i & \text{for TM case, } i \in (0,1). \end{cases} \tag{2.32}$$

In the following, this boundary-value problem will be called normalized grating problem. It is worth noting that equations (2.29) and(2.30) take a simpler form by introducing the total field F:

$$[F]_0 = [F]_1, \tag{2.33}$$

$$\frac{1}{\tau_0}\left[\frac{\mathrm{d}F}{\mathrm{d}n}\right]_0 = -\frac{1}{\tau_1}\left[\frac{\mathrm{d}F}{\mathrm{d}n}\right]_1. \tag{2.34}$$

2.2.5 The special case of the perfectly-conducting grating

The first grating theories were devoted to perfectly conducting gratings. This case is very important since it is realistic for metallic gratings in the microwave domain and far infrared regions. In the visible and infrared regions, it can provide qualitative results. However, in these regions, one must be very cautious. The existence of surface plasmons propagating at the vicinity of the grating surface generates strong resonance phenomena for TM case. **Due to these phenomena, the properties of real metallic gratings and those of perfectly-conducting gratings may completely differ** [2].Moreover, the perfect conductivity model allows one to simplify the grating theory, since the associated boundary-value problems are much simpler.

Basically, the equations associated to the perfect conductivity model are the same as for real metallic or dielectric gratings, except equations (2.4) and (2.11). Let us give a brief explanation to this property. In Maxwell equation (2.4), the right-hand member includes the volume current density $\overrightarrow{j}$ in the metal since this term is proportional to the electric field ($\overrightarrow{j} = \sigma\overrightarrow{E}$, σ being the conductivity of the metal). When the conductivity tends to infinity, the volume current density and the total fields concentrate more and more on the grating surface since the skin depth tends to zero. As a consequence, at the limit when the conductivity tends to infinity, the fields are null in $\mathscr{R}_1$ while the volume current density $\overrightarrow{j}$ becomes a surface current density $\overrightarrow{j_{\mathscr{P}}}$. This surface current density cannot be included in the right-hand member of equation (2.4) since it is a singular distribution (for the specialist of Schwartz distributions [7], it writes $\overrightarrow{j_{\mathscr{P}}}\delta_{\mathscr{P}}$). Finally, equation (2.4) becomes:

$$\nabla\times\overrightarrow{H} = -i\omega\tilde{\varepsilon}\overrightarrow{E} + \overrightarrow{j_{\mathscr{P}}}, \tag{2.35}$$

with $\tilde{\varepsilon}$ being the permittivity of the material. Furthermore, taking into account that the total fields vanish inside $\mathscr{R}_1$, the boundary condition (equation (2.11)) becomes:

$$\overrightarrow{n} \times (\overrightarrow{[H^d]_0} + \overrightarrow{[H^i]_0}) = \overrightarrow{j_{\mathscr{P}}}. \tag{2.36}$$

This equation reduces to a relation between the surface current density on $\mathscr{P}$ and the limit of the magnetic field above $\mathscr{P}$. It does not constitute any more an element of the boundary-value problem.

In conclusion, for perfectly conducting gratings, the fields inside $\mathscr{R}_1$ vanish and, using equations (2.3), (2.4), (2.10), (2.6) and (2.7), the basic vector equations for the field in $\mathscr{R}_0$ can be written:

$$\nabla \times \overrightarrow{E^d} = i\omega\mu_0 \overrightarrow{H^d}, \tag{2.37}$$

$$\nabla \times \overrightarrow{H^d} = -i\omega\varepsilon_0 \overrightarrow{E^d}, \tag{2.38}$$

$$(\overrightarrow{[E^d]_0} + \overrightarrow{[E^i]_0}) \times \overrightarrow{n} = 0. \tag{2.39}$$

Following the same lines as in subsections 2.2.4.1 and 2.2.4.2, the boundary value problems for perfectly conducting gratings are given by:

For TE case:

$$\nabla^2 E_y^d + k_0^2 E_y^d = 0, \tag{2.40}$$

$$\left[E_y^d\right]_0 = -P_y \exp\big(ik_0 x\sin(\theta) - ik_0 z\cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P}, \tag{2.41}$$

$$E_y^d \text{ must satisfy a radiation condition for } z \to +\infty. \tag{2.42}$$

For TM case:

$$\nabla^2 H_y^d + k_0^2 H_y^d = 0, \tag{2.43}$$

$$\left[\frac{\mathrm{d}H_y^d}{\mathrm{d}n}\right]_0 = -iQ_y \overrightarrow{n}.\overrightarrow{k_0^i} \exp\big(ik_0 x\sin(\theta) - ik_0 z\cos(\theta)\big), \quad \text{with } (x,z) \in \mathscr{P}, \tag{2.44}$$

$$H_y^d \text{ must satisfy a radiation condition for } z \to +\infty. \tag{2.45}$$

2.3 Pseudo-periodicity of the field and Rayleigh expansion

This section establishes the most famous property of diffraction gratings: the dispersion of light, which is a consequence of the well known grating formula. In general, this formula is demonstrated using heuristic considerations of physical optics. Here, we propose a rigorous demonstration based on the boundary-value problem stated in subsection 2.2.4.3. First, let us show that the field F^d is pseudo-periodic, i.e. that:

$$F^d(x+d,z) = F^d(x,z) exp\big(ik_0 d\sin(\theta)\big). \tag{2.46}$$

To this aim, we consider the function $G(x,z)$ defined by:

$$G(x,z) = F^d(x+d,z) exp\big(-ik_0 d\sin(\theta)\big). \tag{2.47}$$

The pseudo-periodicity of F^d is proved if we show that $F^d(x,z) = G(x,z)$. Owing to the uniqueness of the solution of the boundary-value problem defined by equations (2.28), (2.29), (2.30) and (2.31), this equation is satisfied if G obeys the same equations. Obviously, G satisfies these equations since d is the grating period. Thus F^d is pseudo-periodic, with coefficient of pseudo-periodicity $k_0 \sin(\theta)$, as well as F^i and F. Notice that in normal incidence ($\theta = 0$), pseudo-periodicity becomes ordinary periodicity, which in that case is a straightforward property since both grating and incident wave are periodic.

Using the pseudo-periodicity, let us show that the field above and below the grating is a sum of plane waves. With this aim, we notice from equation (2.28) that $F^d(x,z)exp\big(-ik_0 x \sin(\theta)\big)$ has a period d and thus can be expanded in a Fourier series:

$$F^d(x,z)\exp\big(-ik_0 x\sin(\theta)\big) = \sum_{n=-\infty}^{+\infty} F_n^d(z)\exp(2i\pi n x/d). \tag{2.48}$$

Multiplying both members of equation (2.48) by $\exp\big(ik_0 x\sin(\theta)\big)$ yields :

$$F^d(x,z) = \sum_{n=-\infty}^{+\infty} F_n^d(z)\exp(i\alpha_n x), \tag{2.49}$$

with:

$$\alpha_n = k_0\sin(\theta) + 2\pi n/d. \tag{2.50}$$

Introducing this expression of $F^d(x,z)$ in Helmholtz equation (2.28), we find :

$$\sum_{n=-\infty}^{+\infty} \big(\mathrm{d}^2F_n^d(z)/\mathrm{d}z^2 + (k^2-\alpha_n^2)F_n^d(z)\big)\exp(i\alpha_n x) = 0, \tag{2.51}$$

and multiplying both members by $\exp\big(-ik_0 x\sin(\theta)\big)$,

$$\sum_{n=-\infty}^{+\infty} \big(\mathrm{d}^2F_n^d(z)/\mathrm{d}z^2 + (k^2-\alpha_n^2)F_n^d(z)\big)\exp(2i\pi n x/d) = 0. \tag{2.52}$$

It seems, at the first glance, that the left-hand member of equation (2.52) is a Fourier series, and thus that the coefficients of this Fourier series vanish. This is not correct. Indeed, we have to bear in mind that k, defined in equation (2.15) is not a constant. As a consequence, if $0 < y < z_M$, a region called intermediate region in the following, k^2 depends on x and the left-hand member of equation (2.52) is not a Fourier series. However, above and below this intermediate region, k^2 is constant and we can write that the Fourier coefficients vanish:

$$\forall n, \qquad \mathrm{d}^2F_n^d(z)/\mathrm{d}z^2 + \gamma_{0,n}^2 F_n^d(z) = 0 \qquad \text{if } y > z_M, \tag{2.53a}$$

$$\forall n, \qquad \mathrm{d}^2F_n^d(z)/\mathrm{d}z^2 + \gamma_{1,n}^2 F_n^d(z) = 0 \qquad \text{if } y < 0, \tag{2.53b}$$

with:

$$\gamma_{i,n} = \sqrt{(k_i^2 - \alpha_n^2)} \qquad i \in (0,1). \tag{2.54}$$

We deduce that:

$$F_n^d(z) = \begin{cases} I_{0,n}\exp(-i\gamma_{0,n}z) + D_{0,n}\exp(+i\gamma_{0,n}z) & \text{if } y > z_M, \\ D_{1,n}\exp(-i\gamma_{1,n}z) + I_{1,n}\exp(+i\gamma_{1,n}z) & \text{if } y < 0, \end{cases} \tag{2.55}$$

and therefore, using equation (2.49),

$$F^d(x,z) = \begin{cases} \sum_{n=-\infty}^{+\infty} \big(I_{0,n}\exp(i\alpha_n x - i\gamma_{0,n} z) + \\ \qquad + D_{0,n}\exp(i\alpha_n x + i\gamma_{0,n} z)\big) & \text{if } z > z_M, \\ \sum_{n=-\infty}^{+\infty} \big(D_{1,n}\exp(i\alpha_n x - i\gamma_{1,n} z) + \\ \qquad + I_{1,n}\exp(i\alpha_n x + i\gamma_{1,n} z)\big) & \text{if } z < 0. \end{cases} \tag{2.56}$$

Let us remark that equation (2.54) does not assign to $\gamma_{i,n}$ a unique value. However, equation (2.56) shows that its determination can be chosen arbitrarily since a sign change does not modify the value of the field, provided that $I_{0,n}$ and $D_{0,n}$ are permuted. The determination of these constants will be given by:

$$\Re(\gamma_{i,n}) + \Im(\gamma_{i,n}) > 0, \qquad i \in (0,1), \tag{2.57}$$

with $\Re(q)$ and $\Im(q)$ denoting the real and imaginary parts of q.

Equation (2.56) shows that the field above and below the intermediate region can be represented by plane wave expansions. The propagation constants of the plane waves along the x and z axes are respectively equal to α_n and $\pm\gamma_{i,n}$. In the physical problem, some of these plane waves must be rejected since they do not obey the radiation condition. This condition entails that $I_{0,n} = I_{1,n} = 0$ since, according to equation (2.57), the associated plane waves propagate towards the grating profile. Finally, equations (2.56), (2.27) and the radiation condition allow us to express the total field by adding the incident field:

$$F(x,z) = \begin{cases} \exp(i\alpha_0 x - i\gamma_{0,0} z) + \\ \qquad + \sum_{n=-\infty}^{+\infty} D_{0,n}\exp(i\alpha_n x + i\gamma_{0,n} z) & \text{if } z > z_M, \\ \sum_{n=-\infty}^{+\infty} D_{1,n}\exp(i\alpha_n x - i\gamma_{1,n} z) & \text{if } z < 0, \end{cases} \tag{2.58}$$

the sums being the expression of the scattered field in both regions. The unknown complex coefficients $D_{0,n}$ and $D_{1,n}$ are the amplitudes of the reflected and transmitted waves respectively.

The conclusion of this subsection is that above and below the intermediate region, the field reflected and transmitted by the grating takes the form of sums of plane waves (Rayleigh expansion [8]), each of them being characterized by its order n.

2.4 Grating formulae

According to equation (2.54), almost all the diffracted plane waves (an infinite number) are evanescent: they propagate along the x axis at the vicinity of the grating profile since they decrease exponentially when $|z| \to +\infty$. For $z \to +\infty$, they correspond to the orders n such that $\alpha_n^2 \geq k_0^2$, thus rendering $\gamma_{0,n} = i\sqrt{(\alpha_n^2 - k_0^2)}$ a purely imaginary number. Only a finite number of them, called z-propagative orders, propagate towards $z = +\infty$, with $\alpha_n^2 \leq k_0^2$, thus $\gamma_{0,n} = \sqrt{(k_0^2 - \alpha_n^2)}$ being real. Let us notice that among these orders, the 0^{th} order is always included, since $\gamma_{0,n} = k_0\cos(\theta)$. It propagates in the direction specularly reflected by the mean plane of the profile, whatever the wavelength may be. In contrast, the other z-propagative orders are dispersive. Indeed, their propagation constants along the x and z axes are equal to α_n and $\gamma_{0,n}$, in such a way that the diffraction angle $\theta_{0,n}$ of one of these waves, measured clockwise from the z axis, can be deduced from $\alpha_n = k_0\sin(\theta_{0,n})$. Using the expression of α_n given by equation (2.50), the angle of diffraction is given by :

$$\boxed{\sin(\theta_{0,n}) = \sin(\theta) + n\frac{2\pi}{k_0 d} = \sin(\theta) + n\frac{\lambda}{d}.} \tag{2.59}$$

This is the famous grating formula, often deduced from heuristic arguments of physical optics.

For the field below the grooves, the wavenumber k_0 is replaced by $k_1 = k_0\nu$. If the grating material is a lossless dielectric, the directions of propagation of the transmitted field obey a grating formula as well. This formula is similar to equation (2.59) but the angles of transmission $\theta_{1,n}$ can be deduced from $\alpha_n = k_0\nu\sin(\theta_{1,n})$, which yields, using a counterclockwise convention :

$$\boxed{\nu\sin(\theta_{0,n}) = \sin(\theta) + n\frac{2\pi}{k_0 d} = \sin(\theta) + n\frac{\lambda}{d}.} \tag{2.60}$$

The 0^{th} order is always included in the z-propagative orders. It propagates in the direction of transmission by an air/dielectric plane interface, whatever the wavelength may be. In contrast, the other z-propagative orders are dispersive. When the grating material is metallic, the transmitted plane waves are absorbed by the metal and the z-propagating orders below the grooves no longer exist.

In conclusion of this section, **the reflected and transmitted fields include, outside the grooves, a finite number of plane waves propagating to infinity with scattering angles given by the grating formulae. All the orders are dispersive, except the 0^{th} orders**. The reflected 0^{th} order takes the specular direction while for a lossless material, the transmitted 0^{th} order takes the direction transmitted by an air/dielectric plane interface. **Consequently, a polychromatic incident plane wave generates in a given order n different from 0 a sum of plane waves scattered in different directions, i.e. a spectrum**. The measurement of the intensity along this spectrum allows one to determine the spectral power of the incident wave. This dispersion phenomenon is the most important property of diffraction gratings. It explains why this optical component has been one of the most valuable tools in the history of Science.

2.5 Analytic properties of gratings

2.5.1 Balance relations

The mathematical balance relations established in this subsection will allow us to demonstrate very important general properties of gratings. These balance relations state mathematical links between characteristics of the field in two regions separated by large distances, without considering the fields in between. They can give a relation between the fields at $z = +\infty$ and the fields on the grating profile, or the fields at $z = -\infty$ and the fields on the grating profile, or the fields at $z = +\infty$ and $z = -\infty$.

2.5.1.1 Lemma 1

We consider two pseudoperiodic functions u and v of the two variables x and z, defined in $\mathscr{R}_0$, which belong to the class G_0 of functions having the following properties:

- They are pseudo-periodic, with the same coefficient of pseudo-periodicity α,
- They are solutions of a Helmholtz equation:

$$\nabla^2 u + k_0^2 u = 0, \tag{2.61a}$$

$$\nabla^2 v + k_0^2 v = 0, \tag{2.61b}$$

with k_0 being real.

- They are bounded for $z \to \infty$,
- They are square integrable in x and locally square integrable in z,
- Their values on $\mathscr{P}$ are square integrable, as well as their normal derivatives.

We introduce the sesquilinear functional defined by:

$$\mathscr{F}_0 = \int_{\mathscr{P}} \left(u \frac{\mathrm{d}\overline{v}}{\mathrm{d}n} - \overline{v} \frac{\mathrm{d}u}{\mathrm{d}n} \right) \mathrm{d}s. \tag{2.62}$$

The symbol $\int_{\mathscr{P}}$ denotes a curvilinear integral on one period of the profile $\mathscr{P}$ of the grating, with ds being the differential of the curvilinear abscissa on $\mathscr{P}$. Obviously, the value in region $\mathscr{R}_0$ of the fields $F(x,z)$, solutions of the four boundary-value problems defined in subsection 2.2.4, belong to G_0, as well as the incident field F^i. It is to be noticed that we do not impose a boundary condition on $\mathscr{P}$ or a radiation condition at infinity, but we still impose that these functions must remain bounded at infinity.

Following the same lines as in section 2.3, it can be shown that above the top of the grooves, u and v can be represented by plane wave expansions, similar to that of equation (2.56): if $z > z_M$,

$$u(x,z) = \sum_{n=-\infty}^{+\infty} [I_{0,n} \exp(i\alpha_n x - i\gamma_{0,n} z) + D_{0,n} \exp(i\alpha_n x + i\gamma_{0,n} z)], \tag{2.63a}$$

$$v(x,z) = \sum_{n=-\infty}^{+\infty} [I'_{0,n} \exp(i\alpha_n x - i\gamma_{0,n} z) + D'_{0,n} \exp(i\alpha_n x + i\gamma_{0,n} z)]. \tag{2.63b}$$

Let us notice that some terms must be eliminated in the Rayleigh expansions. Indeed, the field must remain bounded at infinity. It is not the case for the incident terms of coefficients $I_{0,n}$ and $I'_{0,n}$ unless the corresponding plane waves are z-propagating waves. Thus we define the set U_0 of orders corresponding to z-propagating waves and equations (2.63) become:

$$u(x,z) = \sum_{n \in U_0} I_{0,n} \exp(i\alpha_n x - i\gamma_{0,n} z) + \sum_{n=-\infty}^{+\infty} D_{0,n} \exp(i\alpha_n x + i\gamma_{0,n} z), \tag{2.64a}$$

$$v(x,z) = \sum_{n \in U_0} I'_{0,n} \exp(i\alpha_n x - i\gamma_{0,n} z) + \sum_{n=-\infty}^{+\infty} D'_{0,n} \exp(i\alpha_n x + i\gamma_{0,n} z), \tag{2.64b}$$

$$\begin{aligned} \overline{v(x,z)} = \sum_{n \in U_0} \overline{I'_{0,n}} \exp(-i\alpha_n x + i\gamma_{0,n} z) + \\ + \sum_{n=-\infty}^{+\infty} \overline{D'_{0,n}} \exp(-i\alpha_n x - i\overline{\gamma_{0,n}} z). \end{aligned} \tag{2.64c}$$

Now, we show that $\mathscr{F}_0$ can be expressed as a function of the Rayleigh coefficients $I_{0,n}$, $D_{0,n}$, $I'_{0,n}$ and $D'_{0,n}$. With this aim, we multiply equation (2.61a) by $\overline{v}$, the conjugate of equation (2.61b) by u and we substract the first from the second, which yields:

$$u\nabla^2\overline{v} - \overline{v}\nabla^2 u = 0 \qquad \text{in} \quad \mathscr{R}_0. \tag{2.65}$$

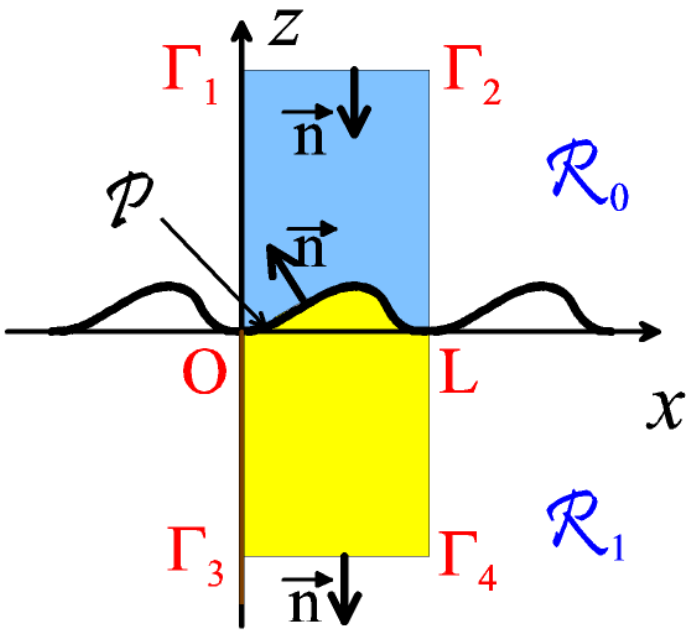


Figure 2.2: Balance relations.

Integrating equation (2.65) in the blue area of figure 2.2 and applying the second Green identity yields:

$$\int_{\Omega_0} \left(u \frac{\mathrm{d}\overline{v}}{\mathrm{dn}} - \overline{v} \frac{\mathrm{d}u}{\mathrm{dn}} \right) \mathrm{d}l = 0, \tag{2.66}$$

with Ω_0 being the boundary of the blue area of figure 2.2 and $\mathrm{d}l$ denoting the differential of the curvilinear abscissa on Ω_0. According to equations (2.64a) and (2.64c), $u\dfrac{\mathrm{d}\overline{v}}{\mathrm{dx}}$ and $\overline{v}\dfrac{\mathrm{d}u}{\mathrm{dx}}$ are periodic. Since the orientations of the normal on verticals $\mathrm{O}\Gamma_1$ and $\mathrm{L}\Gamma_2$ are opposite, the contributions of the integrals on these segments cancel each other. Furthermore, the normal to $\mathrm{O}\Gamma_1$ and $\mathrm{L}\Gamma_2$ is parallel to the z axis and oriented downwords, then equation (2.66) becomes:

$$\int_{\mathscr{P}} \left(u \frac{\mathrm{d}\overline{v}}{\mathrm{dn}} - \overline{v} \frac{\mathrm{d}u}{\mathrm{dn}} \right) \mathrm{d}s = \int_{\Gamma_1\Gamma_2} \left(u \frac{\mathrm{d}\overline{v}}{\mathrm{dz}} - \overline{v} \frac{\mathrm{d}u}{\mathrm{dz}} \right) \mathrm{d}x. \tag{2.67}$$

Introducing in the right-hand member of equation (2.66) the expressions of u and $\overline{v}$ given by equations (2.64a) and (2.64c), separating the terms $n \in U_0$ from the other ones and taking into account that $\int_{x=0}^{d} \exp in\frac{2\pi}{d}x = \delta_{n,0}$, with $\delta_{n,0}$ being the Kronecker symbol, one can obtain, after some cumbersome but not difficult calculations that:

$$\boxed{\int_{\mathscr{P}} \left(u \frac{\mathrm{d}\overline{v}}{\mathrm{dn}} - \overline{v} \frac{\mathrm{d}u}{\mathrm{dn}} \right) \mathrm{d}s = \sum_{n \in U_0} \gamma_{0,n} (I_{0,n} \overline{I'_{0,n}} - D_{0,n} \overline{D'_{0,n}}).} \tag{2.68}$$

2.5.1.2 Lemma 2

In this section, it is supposed that the grating material is lossless, in such a way that plane waves can propagate in $\mathscr{R}_1$. Lemma 2 is similar as lemma 1, but for region $\mathscr{R}_1$. We denote by U_1 the

set of orders corresponding to z-propagating waves in $\mathscr{R}_1$. The expressions of u and v below the x axis are given by:

$$u(x,z) = \sum_{n\in U_1} D_{1,n}\exp(i\alpha_n x - i\gamma_{1,n} z) + \sum_{n=-\infty}^{+\infty} I_{1,n}\exp(i\alpha_n x + i\gamma_{1,n} z), \tag{2.69a}$$

$$v(x,z) = \sum_{n\in U_1} D'_{1,n}\exp(i\alpha_n x - i\gamma_{1,n} z) + \sum_{n=-\infty}^{+\infty} I'_{1,n}\exp(i\alpha_n x + i\gamma_{1,n} z), \tag{2.69b}$$

$$\begin{aligned}\overline{v(x,z)} = \sum_{n\in U_1} \overline{D'_{1,n}}\exp(-i\alpha_n x + i\gamma_{1,n} z) + \\ + \sum_{n=-\infty}^{+\infty} \overline{I'_{1,n}}\exp(-i\alpha_n x - i\overline{\gamma_{1,n}} z).\end{aligned} \tag{2.69c}$$

Following the same lines as in section 2.5.1.1 but for the yellow area of figure 2.2 and noting that the normal is now oriented towards the exterior of the domain, it can be deduced that:

$$\boxed{\int_{\mathscr{P}} \left(u\frac{\mathrm{d}\overline{v}}{\mathrm{d}n} - \overline{v}\frac{\mathrm{d}u}{\mathrm{d}n} \right) \mathrm{d}s = -\sum_{n\in U_1} \gamma_{1,n}(I_{1,n}\overline{I'_{1,n}} - D_{1,n}\overline{D'_{1,n}}).} \tag{2.70}$$

2.5.2 Compatibility between Rayleigh coefficients

In order to state a relation between the Rayleigh coefficients above and below the grating profile, we assume that the functions u and v satisfy the boundary conditions imposed on the total fields by equations (2.33) and (2.34). On the other hand, we do not impose radiation conditions at infinity, but the functions must remain bounded. In other words, u and v can be considered as solutions of the most general grating problem, in which the incident wave is not restricted to a single plane wave, but to the sum of all the plane waves generating diffracted waves in the same directions, with arbitrary amplitudes. It is straightforward to show from equations (2.33) and (2.34) that the left-hand members of equations (2.68) and (2.70) are proportional, then to deduce a relation including the coefficients of the Rayleigh expansions of the field only:

$$\boxed{\begin{aligned}\frac{1}{\tau_0}\sum_{n\in U_0} \gamma_{0,n}(I_{0,n}\overline{I'_{0,n}} - D_{0,n}\overline{D'_{0,n}}) + \\ \frac{1}{\tau_1}\sum_{n\in U_1} \gamma_{1,n}(I_{1,n}\overline{I'_{1,n}} - D_{1,n}\overline{D'_{1,n}}) = 0.\end{aligned}} \tag{2.71}$$

This equation states the most general relation of compatibility between two solutions of the general diffraction grating problem associated to different sets of incident waves. When the grating material is perfectly conducting, it is easy to show that the compatibility equation holds, provided that the sum $n \in U_1$ is cancelled in equation (2.71).

Phenomenological theories of gratings make a wide use of the notion of scattering matrix (or S-matrix). The scattering matrix states the linear relation between the amplitudes of the diffracted and incident waves. We define the column matrix containing the amplitudes of the incident waves. More precisely, we define the normalized amplitudes of the incident and

scattered waves by $\tilde{I}_{0,n} = \sqrt{\gamma_{0,n}} I_{0,n}$, $\tilde{D}_{0,n} = \sqrt{\gamma_{0,n}} D_{0,n}$, $\tilde{I}_{1,n} = \sqrt{\frac{\tau_0}{\tau_1}\gamma_{1,n}} I_{1,n}$, $\tilde{D}_{1,n} = \sqrt{\frac{\tau_0}{\tau_1}\gamma_{1,n}} D_{1,n}$, $n \in (0,1)$, and by definition, the scattering matrix is a square matrix defined by:

$$\mathbb{D} = \mathbb{S}\mathbb{I}, \tag{2.72}$$

with $\mathbb{I}$ being a column vector containing successively all the incident amplitudes $\tilde{I}_{0,n}$ for $n \in U_0$ and all the incident amplitudes $\tilde{I}_{1,n}$ for $n \in U_1$, $\mathbb{D}$ being a column vector containing successively all the diffracted amplitudes $\tilde{D}_{0,n}$, and all the incident amplitudes $\tilde{D}_{1,n}$ for $n \in U_1$. Thus, the order of column matrices $\mathbb{I}$ and $\mathbb{D}$ is the sum$|U_0| + |U_1|$ of the cardinals of U_0 and U_1 Using these notations, equation (2.71) can be expressed in the very simple form:

$$< \mathbb{D}|\mathbb{D}' > = < \mathbb{I}|\mathbb{I}' >, \tag{2.73}$$

the scalar product of two column matrices of order N being defined by:

$$< P|Q > = \sum_{j=1}^{N} P_j \overline{Q_j}. \tag{2.74}$$

Using equation (2.72) to eliminate $\mathbb{D}$ in equation (2.77) yields:

$$< \mathbb{S}\mathbb{I}|\mathbb{S}\mathbb{I}' > = < (\mathbb{S}^*\mathbb{S})\mathbb{I}|\mathbb{I}' > = < \mathbb{I}|\mathbb{I}' >, \tag{2.75}$$

with $\mathbb{S}^*$ being the adjoint matrix of $\mathbb{S}$. Since equation (2.75) must be satisfied for any value of $\mathbb{I}$ and $\mathbb{I}'$, we deduce that:

$$\boxed{\mathbb{S}^*\mathbb{S} = \mathbb{1},} \tag{2.76}$$

with $\mathbb{1}$ being the identity matrix. Equation (2.76) shows that $\mathbb{S}$ is unitary.

2.5.3 Energy balance

The energy balance relation is obtain by taking $u = v$ in equation (2.77), which gives:

$$< \mathbb{D}|\mathbb{D} > = < \mathbb{I}|\mathbb{I} >, \tag{2.77}$$

or equivalently:

$$\|\mathbb{D}\| = \|\mathbb{I}\|. \tag{2.78}$$

Let us show why this equation is known as energy balance relation. To this end, it suffices to use the Poynting theorem and to calculate the flux of the Poynting vector $\overrightarrow{E} \times \overrightarrow{H}$ through the rectangle $\Gamma_1\Gamma_2\Gamma_4\Gamma_3$ of figure 2.2. Since the grating material is lossless, the flux of the Poynting vector through this rectangle (with now the normal oriented toward the exterior, in contrast with figure 2.2) must be null. The contributions of the vertical sides $\Gamma_1\Gamma_3$ and $\Gamma_2\Gamma_4$ cancel each other, thanks to the periodicity of the Poynting vector ($\overrightarrow{H}$ has a coefficient of pseudo-periodicity which is the opposite to that of $\overrightarrow{E}$). At the top of the rectangle, the calculation of the flux of the Poynting vector can be achieved by using the Rayleigh expansion given by equations (2.64). Taking into account that $\int_{x=0}^{d} \exp in\frac{2\pi}{d}x = \delta_n$, elementary calculations show that the contributions to this flux of the different plane waves are decoupled and are proportional to $-\gamma_{0,n}|I_{0,n}|^2$ and $+\gamma_{0,n}|D_{0,n}|^2$. At the bottom of the rectangle, we use the Rayleigh expansion given by equations (2.69). The contributions of the plane waves are decoupled as well and are

proportional to $-\frac{\tau_0}{\tau_1}\gamma_{1,n}|I_{1,n}|^2$ and $+\frac{\tau_0}{\tau_1}\gamma_{1,n}|D_{1,n}|^2$, with the same coefficient of proportionality as the contributions on the top of the rectangle. Therefore, the energy balance can be written:

$$\boxed{\begin{aligned}\sum_{n\in U_0}\gamma_{0,n}|D_{0,n}|^2+\sum_{n\in U_1}\frac{\tau_0}{\tau_1}\gamma_{1,n}|D_{1,n}|^2 = \\ =\sum_{n\in U_0}\gamma_{0,n}|I_{0,n}|^2+\sum_{n\in U_1}\frac{\tau_0}{\tau_1}\gamma_{1,n}|I_{1,n}|^2.\end{aligned}} \tag{2.79}$$

The first and second terms in the left-hand member of equation (2.79) represent the energy diffracted upwards and downwords respectively and the corresponding terms in the right-hand member are the incident energy propagating downwords and upwards respectively.

Coming back to the physical problem where the incident wave is unique and has a unit amplitude (see equation (2.26)), equation (2.79) becomes:

$$\boxed{\sum_{n\in U_0}\gamma_{0,n}|D_{0,n}|^2+\sum_{n\in U_1}\frac{\tau_0}{\tau_1}\gamma_{1,n}|D_{1,n}|^2=\gamma_{0,0},} \tag{2.80}$$

the right-hand member representing the incident energy. In that case, the efficiency $\rho_{i,n}$, $i\in(0,1)$ is defined as the ratio of the energy diffracted in a given order over the incident energy. Using equation (2.79) yields:

$$\rho_{i,n}=\begin{cases}\dfrac{\gamma_{0,n}}{\gamma_{0,0}}|D_{0,n}|^2 & \text{if } i=0,\\ \dfrac{\tau_0}{\tau_1}\dfrac{\gamma_{1,n}}{\gamma_{0,0}}|D_{1,n}|^2 & \text{if } i=1,\end{cases} \tag{2.81}$$

and the energy balance can be written:

$$\sum_{n\in U_0}\rho_{0,n}+\sum_{n\in U_1}\rho_{1,n}=1. \tag{2.82}$$

The sum of efficiencies is equal to unity. When the grating is perfectly conducting, it is easy to show that the energy balance still holds, provided that the sum $n\in U_1$ is cancelled in equations (2.79), (2.80) and (2.82). When the grating material is lossy, the sum $n\in U_1$ must be cancelled as well and one can show that equation (2.82) becomes:

$$\sum_{n\in U_0}\rho_{0,n}<1. \tag{2.83}$$

The sum of reflected efficiencies is smaller than one, a rather intuitive result if we bear in mind that a part of the incident energy is dissipated in the grating material.

2.5.4 Reciprocity

In order to demonstrate the well known reciprocity relation, we consider a function u, sum of the solution of the normalized grating problem (see equations(2.28), (2.29), (2.30) and (2.31)) and of the corresponding incident field (in other words, u is the total field). In order to define v, we consider the p^{th} order of diffraction ($p\in U_0$) in $\mathscr{R}_0$, with diffraction angle $\theta_{0,p}$.

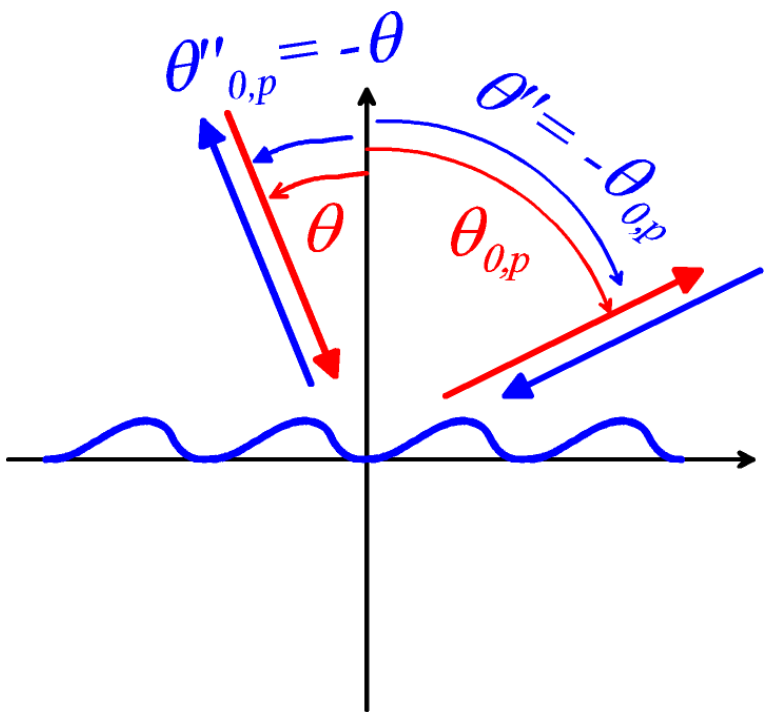


Figure 2.3: The reciprocity theorem: The efficiency in the p^{th} order is the same in the two cases symbolized by red and blue arrows.

Then, we consider a second problem, but with angle of incidence $\theta'' = -\theta_{0,p}$, as shown[2] in figure 2.3. The incident wave in this second case has a direction of propagation which is just the opposite of that of the p^{th} diffracted order in the first case and straightforward calculations show that **the corresponding p^{th} order in $\mathscr{R}_0$ has a direction of propagation which is the opposite of that of the incident wave in the first case**, which entails $\theta''_{0,p} = -\theta$. This geometrical property is known in optics as the reversion theorem. The constants of propagation of the p^{th} diffracted order in this second case are given by $\alpha''_p = -\alpha_0$ and $\gamma''_{0,p} = \gamma_{0,0}$ and more generally, the constants of propagation of an arbitrary n^{th} diffracted order in this second case are given by $\alpha''_n = -\alpha_{p-n}$ and $\gamma''_{0,n} = \gamma_{0,p-n}$. Thus v'' can be written:

$$v''(x,z) = \exp(-i\alpha_p x - i\gamma_{0,p} z) + \sum_{n=-\infty}^{+\infty} D''_{0,n} \exp(-i\alpha_{p-n} x + i\gamma_{0,p-n} z). \tag{2.84}$$

Functions u and v'' do not satisfy the conditions of the equation of compatibility (equation (2.71)) since they have not the same pseudo-periodicity. It is not so for u and the function $v = \overline{v''}$ which is given by:

$$v(x,z) = \exp(i\alpha_p x + i\gamma_{0,p} z) + \sum_{n=-\infty}^{+\infty} \overline{D''_{0,n}} \exp(i\alpha_{p-n} x - i\overline{\gamma_{0,p-n}} z). \tag{2.85}$$

Identifying the incident and diffracted waves in equation (2.85) yields:

$$I'_{0,n} = \overline{D''_{0,p-n}}, \tag{2.86a}$$

$$D'_{0,n} = \delta_{n-p}, \tag{2.86b}$$

[2]It must be remembered that the conventions for the measurements of the angles of incidence and diffraction in $\mathscr{R}_0$ are opposite

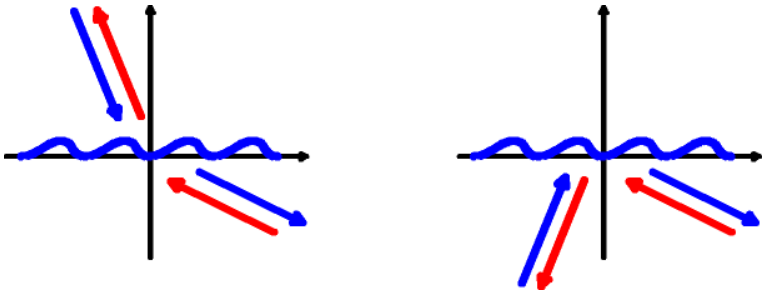


Figure 2.4: Other reciprocity relations: The efficiency is the same in the two cases symbolized by red and blue arrows.

and from equation (2.71), it turns out that:

$$\boxed{\gamma''_{0,p} D''_{0,p} = \gamma_{0,p} D_{0,p}.} \tag{2.87}$$

This is the reciprocity theorem: the products of the amplitudes of the plane waves represented in figure 2.3 by their propagation constants along the z axis is invariant. In order to state the reciprocity theorem in a form which is most widespread, we take the modulus square of both members of equation (2.87):

$$\gamma''_{0,p}{}^2 |D''_{0,p}|^2 = \gamma_{0,p}{}^2 |D_{0,p}|^2. \tag{2.88}$$

Writing equation (2.88) in the form:

$$\frac{\gamma''_{0,p}}{\gamma_{0,p}} |D''_{0,p}|^2 = \frac{\gamma_{0,p}}{\gamma''_{0,p}} |D_{0,p}|^2, \tag{2.89}$$

and bearing in mind that $\gamma_{0,p} = \gamma''_{0,0}$ and $\gamma''_{0,p} = \gamma_{0,0}$, and using the definition of the efficiencies given in equation (2.81), equation (2.89) yields:

$$\boxed{\rho''_{0,p} = \rho_{0,p}.} \tag{2.90}$$

The efficiency is invariant.

Figure 2.4 illustrates two other cases where the reciprocity theorem applies. These properties can be demonstrated by following the same lines as in the first part of this section. It is important to notice that the reciprocity theorem illustrated in figure 2.3 holds for lossy materials [9]. More surprisingly, the theorem can be generalized to evanescent waves [10].

2.5.5 Uniqueness of the solution of the grating problem

If two different solutions of the normalized grating problem exist, their difference $w(x,z)$ does not include any incident wave. We will show that such a field vanish. We assume here that the grating material is lossless. First, using the compatibility equation (2.71) with $u = v = w$, it emerges that:

$$\frac{1}{\tau_0} \sum_{n \in U_0} \gamma_{0,n} |D_{0,n}|^2 + \frac{1}{\tau_1} \sum_{n \in U_1} \gamma_{1,n} |D_{1,n}|^2 = 0. \tag{2.91}$$

Since τ_0, τ_1, $\gamma_{0,n}$ and $\gamma_{1,n}$are positive, equation (2.91) implies that $D_{0,n} = D_{1,n} = 0$. This is an important result since it means that if w exists, it has no effect on the far field: the solution in the far field is unique. However, it could exist a function w localized at the vicinity of the grating profile and tending to zero exponentially at infinity. The interested reader can find a complete and not straightforward demonstration of the uniqueness in [1], at least for the TE case.

2.5.6 Analytic properties of crossed gratings

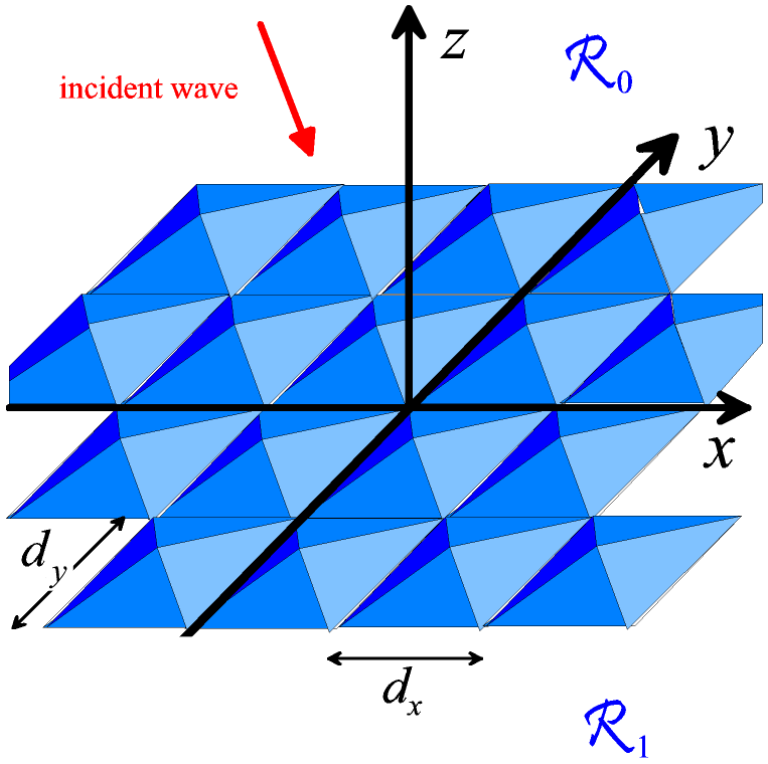


Figure 2.5: A crossed grating with periods d_x and d_z on the x and z axes.

Now, we consider the diffraction problem schematized in figure 2.5. An incident wave of wavevector k_0 is incident on a doubly-periodic structure separating air (region $\mathscr{R}_0$) from a grating material (region $\mathscr{R}_1$). We use all the notations defined in the preceding sections to characterize the materials. The direction of incidence is specified by the polar angles Φ and Ψ. In order to define the polarization of the incident field, we construct the circle MNM'N' in the plane perpendicular to k_0, with the continuation of NN' intersecting the z axis and MM' being perpendicular to NN'. The polarization angle δ is the angle between M'M and the direction of the incident electric field $\overrightarrow{P}$. With these notations, the incident electric field is given by:

$$\overrightarrow{E^i} = \overrightarrow{P}\exp\left(i\alpha x + i\beta y - i\gamma z\right), \tag{2.92}$$

with $\alpha = k_0 \sin\Phi\cos\Psi$, $\beta = k_0 \sin\Phi\sin\Psi$ and $\gamma = k_0\cos\Phi$. The projection of $\overrightarrow{P}$ on M'M is called transverse component of $\overrightarrow{P}$ and denoted by P^t. Its projection on N'N is called longitudinal (in plane) component and denoted bzP^l, in such a way that $\overrightarrow{P} = P^t\dfrac{\overrightarrow{MM'}}{MM'} + P^l\dfrac{\overrightarrow{NN'}}{NN'}$.

As in the case of classical gratings, it is possible to show that above the top of the grating ($z > z_M$), the field can be expanded in the form of a sum of plane waves:

$$\overrightarrow{E}(x,z) = \begin{cases} \Sigma_{n=-\infty}^{+\infty}\Sigma_{m=-\infty}^{+\infty}\left(\overrightarrow{I_{0,n,m}}\exp(i\alpha_n x + i\beta_m y - i\gamma_{0,n,m}z) + \right. \\ \quad \left. + \overrightarrow{D_{0,n,m}}\exp(i\alpha_n x + i\beta_m y + i\gamma_{0,n,m}z)\right), \qquad \text{if } z > z_M, \\ \Sigma_{n=-\infty}^{+\infty}\Sigma_{m=-\infty}^{+\infty}\left(\overrightarrow{D_{1,n,m}}\exp(i\alpha_n x + i\beta_m y - i\gamma_{1,n,m}z) + \right. \\ \quad \left. + \overrightarrow{I_{1,n,m}}\exp(i\alpha_n x + i\beta_m y + i\gamma_{1,n,m}z)\right) \qquad \text{if } z < 0. \end{cases} \tag{2.93}$$

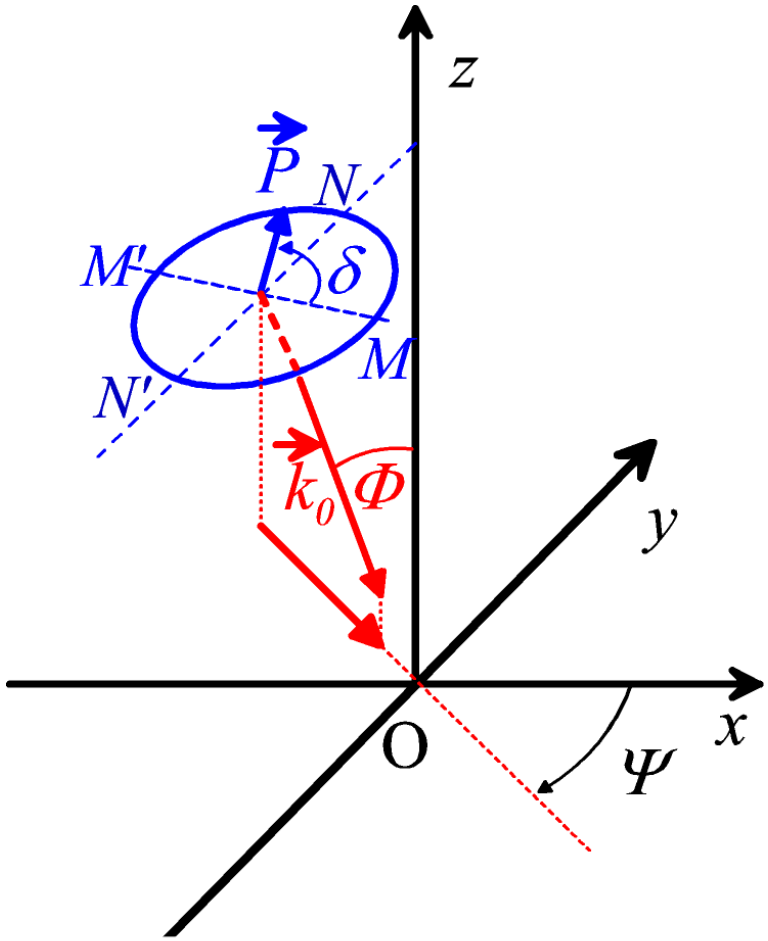


Figure 2.6: Notations for the incident field.

The wavevectors of all these plane waves must be orthogonal to their vector amplitudes. As for the incident wave, we can define the transverse and longitudinal components of the vector amplitudes of the plane waves, the transverse component (for example $D^t_{0,n,m}$) being orthogonal to the z axis in the plane perpendicular to the wavevector $(\alpha_n, \gamma_{0,n,m}, \beta_m)$ and the longitudinal (for example $D^l_{0,n,m}$) its component in the orthogonal direction of the same plane.

Using the Poynting theorem, it can be shown, as in section 2.5.3, that the efficiencies in the z-propagating orders are given by:

$$\rho_{i,n,m} = \begin{cases} \dfrac{\gamma_{0,n,m}}{\gamma_{0,0}} |\overrightarrow{D}_{0,n,m}|^2 & \text{if } i = 0, \\ \dfrac{\gamma_{1,n,m}}{\gamma_{0,0}} \left(\dfrac{1}{\nu^2} |D^l_{1,n,m}|^2 + |D^t_{1,n,m}|^2 \right) & \text{if } i = 1. \end{cases} \tag{2.94}$$

Of course, the line associated to $i = 1$ in equation (2.94) must be cancelled if the grating material is lossy.

We define, as for classical gratings, the sets U_0 and U_1 of z-propagating orders in $\mathscr{R}_0$ and $\mathscr{R}_1$ respectively and, when the grating material is lossless, the energy balance can be written:

$$\boxed{\sum_{(n,m)\in U_0} \rho_{0,n,m} + \sum_{(n,m)\in U_1} \rho_{1,n,m} = 1.} \tag{2.95}$$

We will not demonstrate the reciprocity theorem, the interested reader can find the proof in [1]. This theorem, in the case of an order (p,q) propagating in $\mathscr{R}_0$ can be expressed in the following form:

$$\boxed{\gamma \overrightarrow{P}.\overrightarrow{D'}_{0,p,q} = \gamma' \overrightarrow{P'}.\overrightarrow{D}_{0,p,q}.} \tag{2.96}$$

In the first case, the incident electric field with vector amplitude $\overrightarrow{P}$ and propagation constant along the z axis $-\gamma$ generates in $\mathscr{R}_0$ in the (p,q) order, with $(p,q) \in U_0$, a plane wave of vector amplitude $\overrightarrow{D}_{0,p,q}$ and propagation constant along the z axis $\gamma_{0,p,q}$. In the second case, we consider an incident wave which propagates in the direction which is just the opposite to that of the (p,q) order in the first case. Thus its constant of propagation along the z axis is $-\gamma' = -\gamma_{0,p,q}$. The vector amplitude of this incident wave is equal to $\overrightarrow{P'}$. It can be shown that in this second case, the (p,q) order takes the direction which is the opposite of that of the incident wave in the first case and its vector amplitude is equal to $\overrightarrow{D'}_{0,p,q}$. **Thus, equation (2.96) can be expressed in the following form: the scalar product of the vector amplitudes of the incident and diffracted waves propagating in the opposite directions, multiplied by the propagation constant of the incident wave along the z axis, is constant**. It can be shown that this relation entails the **reciprocity in natural light for the efficiencies**:

$$\boxed{< \rho_{0,p,q} > = < \rho'_{0,p,q} >,} \tag{2.97}$$

with $< \rho_{0,p,q} >$ being the average between the efficiencies in both cases of polarization ($\delta = 0$ and $\delta = \frac{\pi}{2}$).

2.6 Conclusion

We have established the mathematical bases of grating theories: the boundary-value problems. Most of the formalisms used for solving the grating problems numerically start from these boundary-value problems, for example the integral theory [1,2].Other theories use some conditions of these problems but deal directly with Maxwell equations, for example the RCWA method [5].

Without any doubt, the boundary-value problems are necessary to demonstrate the analytic properties of gratings. Very often, these properties are ignored or neglected. However, properties like energy balance or reciprocity are needed for a full understanding of the puzzling properties of this crucial component of optics and nanophotonics. These analytic properties are also widely used to check new grating softwares. However, they are not more than casting out nines. They can show that a software fails if they are nor satisfied on its numerical results. It must be emphasized that they can never prove its validity if they are satisfied.

Some important analytic properties of gratings have not been mentioned in this chapter. It is the case for example for the Marechal and Stroke theorem, the only grating property which allows one to know the field diffracted by a grating without any calculation. This theorem, which is restricted to perfecly conducting echelette gratings used for TM polarization in very special conditions will be given in the chapter devoted to the applications of grating properties.

References:

[1] R. Petit, Ed.: Electromagnetic theory of gratings. *Topics in current physics*, (Springer-Verlag, 1980) .

[2] D. Maystre: Rigorous vector theories of diffraction gratings, In:*Progress in Optics **21***, ed. by E. Wolf (North-Holland) pp. 1-67 (1984) .

[3] M. Nevière, and E. Popov: Light Propagation in Periodic Media: Differential Theory and Design, (Marcel Dekker, 2003) .

[4] L. Li, J. Chandezon, J., G. Granet, and J.-P. Plumet : Rigorous and Efficient Grating-Analysis Method Made Easy for Optical Engineers. *Applied Optics* **38**, 304-313 (1999) .

[5] M. G. Moharam, and T. K. Gaylord : Rigorous coupled-wave analysis of metallic surface-relief gratings. *J. Opt. Soc. Am.* **3**, 1780-1787 (1986) .

[6] L. C. Botten, M. S.Craig, R.C. McPhedran, J. L. Adams, and J. R. Andrewartha : The dielectric lamellar diffraction grating. *Optica Acta* **28**, 413 – 428 (1981) .

[7] L. Schwartz: Mathematics for Physical sciences, (Addison-Wesley, London, 1967) .

[8] Rayleigh, Lord: On the dynamical theory of gratings, *Proc. Royal Soc. A*, **79** pp. 399-416 (1907)

[9] D. Maystre, and R. C. McPhedran: Le théorème de réciprocité pour les réseaux de conductivité finie: démonstration et applications. *Optics Commun.* **12**, 164-167 (1974) .

[10] R. Carminati, M. Nieto-Vesperinas, and J.-J. Greffet: Reciprocity of evanescent electromagnetic waves. *J. Opt. Soc. Am. A* **15**, 706-712 (1998) .

Chapter 3:
Spectral Methods for Gratings

John A. deSanto

Table of Contents:

Chapter 3

Spectral Methods for Gratings

John A. DeSanto

Professor Emeritus, Department of Physics
Colorado School of Mines
Golden, CO 80401, USA
jdesanto@mines.edu, bajdesanto@mac.com

Abstract

We present a unified formal treatment of spectral methods applied to scattering from penetrable gratings for both acoustic (scalar) and electromagnetic (vector) problems. These are derived from coordinate space representations for both acoustic problems for a one-dimensional grating, and full electromagnetic problems for a two-dimensional grating. The coordinate space representations are also derived here. By unified we mean that the electromagnetic results use a scalar analogy, in that the boundary unknowns are the electric field and its normal derivative.

In coordinate space, the kernels of either the integral representations or integral equations have two variables, the first relating to the field coordinate (which, for an integral equation, has been evaluated on the surface), and the second evaluated on the surface as part of the surface integration. We refer to this procedure as coordinate in both variables or simply a *CC* method. Partial spectral results involve a spectral replacement of the first coordinate variable, and we refer to these methods as *SC*. Full spectral methods involve an additional replacement of the second (always surface) coordinate variable by a spectral one and we refer to these as *SS* methods, or in the case of a conjugate Rayleigh basis as SS^*.

For both scalar and electromagnetic cases, the partial spectral results are derived without the use of Green's functions. Instead we use plane wave states in Green's theorem. The partial spectral results are also used to generate surface inversion methods involving perturbation theory and the Kirchhoff approximation. For the full spectral scalar case three spectral expansions are considered, a physical optics modified Fourier expansion, a Floquet-Fourier expansion, and an expansion in conjugate Rayleigh basis functions. All are Floquet- or quasi-periodic. For the full spectral electromagnetic case only the conjugate Rayleigh basis expansion is presented.

3.1 Introduction

In this paper we present formal equations in spectral space to describe the scattering from periodic surfaces or gratings. We do this both for the acoustic case in one dimension for the direct scattering problem (Secs.4,5,7) and the inverse problem(Sec.6), and in two dimensions for the general electromagnetic problem (Secs.9,10,11). In order to do this and justify the validity of the spectral representations in various regions, we derive in each case the coordinate-space representations for the scattering (Secs.4,9). This includes the two- and three-dimensional periodic Green's function (Sec.3) necessary to describe the coordinate-space scattering, and the development of plane wave expansions in periodic media (Sec.2) which are used throughout the paper.

For the partial spectral-space equations it is not necessary to use the Green's function. Instead we describe a method using plane waves and Green's theorem in the periodic cell of the surface to derive the spectral equations directly and simply. One can describe the coordinate-space integral representations or integral equations to solve the boundary unknowns as equations in two coordinate spaces, the space of the source or integrated coordinate, and the space of the field or exterior coordinate. For clarity, we refer to this as a coordinate-coordinate (CC) representation. Spectral can then refer to one or both of these coordinate spaces transformed, a partial spectral space when one is transformed (Secs.5,10) where we use first the transform of the exterior coordinate (following from the plane waves and Green's theorem) so we are in a spectral-coordinate (SC) representation which we treat extensively, including its use in the inverse problem, the problem of surface reconstruction from (known) scattered field data. A second version (CS) is referred to[93] but not extensively discussed. Using SC we can then represent the integrated coordinate in spectral space (SS) where we reference extensive computational results[36,37,38,39]. What has become of interest lately[3] is the transform of the coordinate-space of integration into the conjugate spectral space $S^\star$, and we describe this $SS^\star$ method extensively (Secs.8,11) for acoustic and electromagnetic results respectively. Formally this is equivalent to a dual least squares method.

It is important to define what we mean by the word "spectral"[28]. In a general context, this could mean just Fourier space, and boundary function expansions in pure Fourier series. However, in our context, this is not correct. The reason is that the boundary fields are limits of Floquet- or quasi-periodic functions and must themselves be Floquet- or quasi-periodic. This reflects the nature of the incident field being, in general, incident off the normal. The periodic surface has right-left symmetry, but the boundary value problem does not (unless the field is incident normally). By "spectral" we thus mean here three kinds of expansions, a physical-optics modified Fourier expansion (Sec.7) where a single plane wave modulates the Fourier series, an expansion which we term Floquet-Fourier which preserves the quasi-periodicity but without the modulating plane wave (Sec.7), and an expansion in conjugate plane wave states on the boundary (Secs.8,11) where plane waves modulate each term in the expansion. These latter are Rayleigh or Bloch wave type expansions and are useful because, at least in some degenerate cases, they lead to self-adjoint problems. All expansions are quasi-periodic. Relation of the results to Rayleigh and Waterman expansions is discussed.

Using the partial spectral methods developed in Sec.5 for the direct scattering problem, we discuss in Sec.6 how they can be used to find the periodic surface profile from the incident and scattered fields. Two methods are presented, one based on perturbation theory and the other on the Kirchhoff approximation, both for the scalar case. Both surface inversion methods were initially applied to truncated random surfaces[111,112] with good results, and the methodology is

here applied to gratings.

The electromagnetic equations we present are not done in the conventional formalism[61] using boundary currents, but are based on earlier work of ours[35] which rely on a scalar analogue of the electromagnetic problem, so the boundary unknowns we use are the electric field and its normal derivative. The resulting equations become a direct scalar analogue in terms of different boundary unknowns for the electromagnetic problem. We also do not discuss the computational solutions of the equations but rely on references where available. Some related equations have been solved, and we point out the references where appropriate, but a full discussion of computational issues would require a separate paper.

A very large number of references are cited in the text. Many of these references, some not specifically attuned to spectral methods, nevertheless contain spectral components in the development or in reference to expansions, in particular to the Rayleigh expansion[11,12,14,48,50,54,56,57,64,66,70,76,80,81,82,94,99,100,104], the Waterman expansion[105], both[4,5,23,24,109], or a combination of the two expansions[62]. The gratings we consider are infinite, although spectral methods have been applied to finite gratings also[84], and our gratings are purely deterministic although random gratings have also been considered[83]. Newer results on gratings apply surface integral methods to periodic nanostructures[43], indicating the generality of the methods we describe. We are mainly interested in scattering methods which produce the scattered and transmitted fields and their use in inversion, although others prefer to consider the dispersion relation for surface plasmons and polaritons propagating along the grating[46,47,63]. Other rigorous theoretical and computational developments are also available[2,6,7,8,16,44,45,49,67,68,77,78,79,87,92,96,113], as well as approximations[65,110], applications[9,13,42,51,71,75], and other methodology[1,85].

There are very many papers (our many references do not scratch the surface), reviews[10,40,69,88,89,91,98] and books[52,103,108] on scattering from gratings, not the least of them being the important book edited by Petit[90] in honor of which this paper is contributed.

3.2 Plane Waves in Periodic Media

In two dimensions $\vec{x} = (x,z)$, we write a plane wave as

$$\phi(\vec{x}) = \exp[ik(\alpha_0 x + \gamma_0 z)], \tag{3.1}$$

where $\alpha_0 = \sin(\theta)$, $\gamma_0 = \cos(\theta) = \sqrt{1-\alpha_0^2}$, and θ is measured clockwise from the positive z axis. With the time convention $\exp(-i\omega t)$, where ω is circular frequency, this is thus an up-going plane wave, and satisfies the two-dimensional Helmholtz equation

$$(\nabla_2^2 + k^2)\phi(\vec{x}) = 0, \tag{3.2}$$

where k is the wavenumber (here considered to be strictly real). On a one-dimensional surface $z = h(x)$ we have

$$\phi(\vec{x}_h) = \exp[ik(\alpha_0 x + \gamma_0 h(x))], \tag{3.3}$$

where $\vec{x}_h = (x, h(x))$. This is referred to as a Rayleigh function. For one-dimensional periodic media (here the surface), $h(x+L) = h(x)$, where L is the period, we have the relation

$$\phi(x+L, h(x+L)) = \exp(ik\alpha_0 L)\phi(x, h(x)). \tag{3.4}$$

The same result is true even off the surface, i.e.

$$\phi(x+L, z) = \exp(ik\alpha_0 L)\phi(x, z). \tag{3.5}$$

These results are referred to as Floquet- or quasi-periodicity. The field scattered from this periodic surface, ψ^{sc}, satisfies the same Helmholtz equation, and is also quasi-periodic because the ratio $\frac{\psi^{sc}}{\phi}$ is periodic, i.e.

$$\frac{\psi^{sc}(x+L,z)}{\phi(x+L,z)} = \frac{\psi^{sc}(x,z)}{\phi(x,z)}, \tag{3.6}$$

so that

$$\psi^{sc}(x+L,z) = \exp(ik\alpha_0 L)\psi^{sc}(x,z), \tag{3.7}$$

and the same is true on the surface $z = h(x)$. In general for any field function ψ satisfying (3.2) and shifted an integer n number of periods we have

$$\psi(x+nL,z) = \exp(ik\alpha_0 nL)\psi(x,z). \tag{3.8}$$

The same is of course true on the surface $z = h(x)$.

In three dimensions any field function which satisfies the three-dimensional Helmholtz equation

$$(\nabla_3^2 + k^2)\psi(\vec{x}) = 0, \tag{3.9}$$

where $\vec{x} = (x,y,z)$, and is quasi-periodic in x with period L_1 and in y with period L_2 satisfies

$$\psi(x+n_1L_1, y+n_2L_2, z) = \exp[ik(\alpha_0 n_1 L_1 + \beta_0 n_2 L_2)]\psi(x,y,z), \tag{3.10}$$

where $\alpha_0 = \sin(\theta)\cos(\varphi)$ and $\beta_0 = \sin(\theta)\sin(\varphi)$ with θ the polar angle, φ the azimuthal angle, and n_1 and n_2 are integers. In three dimensions the up-going plane wave is now written as

$$\phi(\vec{x}) = \exp[ik(\alpha_0 x + \beta_0 y + \gamma_0 z)], \tag{3.11}$$

where $\gamma_0 = \sqrt{1-\alpha_0^2-\beta_0^2}$. Although we use some of the same notation in both two and three dimensions the interpretation will be clear from the context.

We consider the periodic surface $z = h(x)$ in one dimension and $z = h(x,y) = h(\vec{x}_\perp)$ in two dimensions which separates two media with wavenumbers k_1 for $z > h$ (the upper region 1) and k_2 for $z < h$ (the lower region 2). Notationally, subscripts are used to identify the region, e.g. ϕ becomes ϕ_1 or ϕ_2, γ_0 becomes γ_{10} or γ_{20}, etc. The paper has a lot of notation, and we have tried to keep it as clear as possible.

3.3 Green's Functions in Periodic Media

In two dimensions the free-space Green's function is

$$G^{(2)}(\vec{x}',\vec{x}) = \frac{i}{4}H_0^{(1)}(k_0|\vec{x}'-\vec{x}|), \tag{3.12}$$

where $H_0^{(1)}$ is the Hankel function, and k_0 is a generic wave number. It satisfies the equation

$$(\nabla_2^2 + k_0^2)G^{(2)}(\vec{x}',\vec{x}) = -\delta(\vec{x}'-\vec{x}). \tag{3.13}$$

Its representation as a Fourier transform is

$$G^{(2)}(\vec{x}',\vec{x}) = \frac{1}{(2\pi)^2}\iint \frac{\exp[ik_x(x'-x)+ik_z(z'-z)]}{k^2-k_{0+}^2}dk_x dk_z, \tag{3.14}$$

where the integrals run from $-\infty$ to ∞. We have given k_0 a small positive imaginary part to define the integral, and $k^2 = k_x^2 + k_z^2$. If we choose a specific direction, here the fixed direction z, we can evaluate the k_z integration using complex variables. The result is the Weyl representation[34] for $G^{(2)}$

$$G^{(2)}(\vec{x}',\vec{x}) = \frac{i\pi}{(2\pi)^2}\int_{-\infty}^{\infty}\frac{\exp[ik_x(x'-x)+iK_0|z'-z|]}{K_0}dk_x, \tag{3.15}$$

where $K_0 = \sqrt{k_0^2 - k_x^2}$ for $k_0^2 > k_x^2$, and $= i\sqrt{k_x^2 - k_0^2}$ for $k_x^2 > k_0^2$.

We have two regions. In region 1, we let $k_0 = k_1$ in (3.15), scale the integral using $k_x = k_1\alpha$, and we get the Green's function for region 1 (subscript) in (2)-dimensions (superscript)

$$G_1^{(2)}(\vec{x}',\vec{x}) = \frac{i\pi}{(2\pi)^2}\int_{-\infty}^{\infty}\frac{\exp[ik_1(\alpha(x'-x)+\gamma_1(\alpha)|z'-z|)]}{\gamma_1(\alpha)}d\alpha, \tag{3.16}$$

where

$$\begin{aligned}\gamma_1(\alpha) &= \sqrt{1-\alpha^2}, \quad (\alpha^2 < 1) &(3.17)\\ &= +i\sqrt{\alpha^2-1}, \quad (\alpha^2 > 1). &(3.18)\end{aligned}$$

In region 2, let $k_0 = k_2$ in (3.15) and scale using $k_x = k_1\alpha$ (the same scaling as in region 1) to get the Green's function in region 2 (subscript) in (2)-dimensions (superscript)

$$G_2^{(2)}(\vec{x}',\vec{x}) = \frac{i\pi}{(2\pi)^2}\int_{-\infty}^{\infty}\frac{\exp[ik_1(\alpha(x'-x)+\gamma_2(\alpha)|z'-z|)]}{\gamma_2(\alpha)}d\alpha, \tag{3.19}$$

where

$$\begin{aligned}\gamma_2(\alpha) &= \sqrt{K^2-\alpha^2}, \quad (\alpha^2 < K^2) &(3.20)\\ &= +i\sqrt{\alpha^2-K^2}, \quad (\alpha^2 > K^2), &(3.21)\end{aligned}$$

and $K = k_2/k_1$, the ratio of wavenumbers. The same scaling in both regions can be thought of as simply a result of Snell's Law since the x-components of the phases of both functions must match at a flat interface.

The Green's functions above are for an infinite space or, in our case, an infinite surface. To find the periodic Green's function for a single cell of the surface we use the single or double layer potentials which occur in Sec.4. For example, define the single layer potential on an infinite surface $h(x)$ as

$$(S\psi)(\vec{x}_h') = \int_{-\infty}^{\infty} G^{(2)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dx, \tag{3.22}$$

where ψ is any field function (here it is the normal derivative). The result can be written as a sum over periodic cells

$$(S\psi)(\vec{x}_h') = \sum_{n=-\infty}^{\infty} I_n(x'), \tag{3.23}$$

where

$$I_n(x') = \int_{(2n-1)L/2}^{(2n+1)L/2} G^{(2)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dx. \tag{3.24}$$

Use (3.16) or (3.19) in (3.24), shift the integration by defining $x'' = x - nL$, and use the Floquet property of the field function to rewrite (3.22) as

$$(S\psi)(\vec{x}_h') = \int_{-L/2}^{L/2} G^{(2p)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dx, \tag{3.25}$$

where the two-dimensional periodic Green's function ($(2p)$-superscript) is given by

$$G^{(2p)}(\vec{x}_h',\vec{x}_h) = \frac{i\pi}{(2\pi)^2}\int_{-\infty}^{\infty}\frac{\exp[ik_1(\alpha(x'-x)+\gamma(\alpha)|h(x')-h(x)|)]}{\gamma(\alpha)}S(\alpha)d\alpha, \tag{3.26}$$

where the sum S is given by

$$S(\alpha) = \sum_{n=-\infty}^{\infty}\exp[ink_1L(\alpha_0-\alpha)], \tag{3.27}$$

and can be evaluated using the Poisson sum[95] to be

$$S(\alpha) = \frac{2\pi}{ik_1}\sum_{j=-\infty}^{\infty}\delta(\alpha-\alpha_j), \tag{3.28}$$

where δ represents the delta function and $\alpha_j = \alpha_0 + j\lambda/L$ is the grating equation. The result substituted in (3.26) yields the periodic Green's function for region 1

$$G_1^{(2p)}(\vec{x}_h',\vec{x}_h) = \frac{i}{2k_1L}\sum_{j=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\gamma_{1j}|h(x')-h(x)|)]}{\gamma_{1j}}, \tag{3.29}$$

where

$$\gamma_{1j} = \sqrt{1-\alpha_j^2}, \quad (\alpha_j^2<1), \tag{3.30}$$

$$= +i\sqrt{\alpha_j^2-1}, \quad (\alpha_j^2>1), \tag{3.31}$$

and the periodic Green's function for region 2

$$G_2^{(2p)}(\vec{x}_h',\vec{x}_h) = \frac{i}{2k_1L}\sum_{j=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\gamma_{2j}|h(x')-h(x)|)]}{\gamma_{2j}}, \tag{3.32}$$

where

$$\gamma_{2j} = \sqrt{K^2-\alpha_j^2}, \quad (\alpha_j^2<K^2) \tag{3.33}$$

$$= +i\sqrt{\alpha_j^2-K^2}, \quad (\alpha_j^2>K^2). \tag{3.34}$$

We have listed the Green's functions of both regions to stress the exterior scaling k_1 for both. The result is the residual of the k_1 in the phase of both terms, Snell's law, and the same Poisson sum.

The Green's functions satisfy the differential equations

$$(\nabla_2^2+k_l^2)G_l^{(2p)}(\vec{x}',\vec{x}) = -\sum_{n=-\infty}^{\infty}\delta(x'-x_n)\delta(z'-z), \tag{3.35}$$

where $x_n = x+nL$ and $l = 1,2$. The periodic Green's function can also be written as a phased array of Hankel functions, e.g. for region 1

$$G_1^{(2p)}(\vec{x}',\vec{x}) = \frac{i}{4}\sum_{n=-\infty}^{\infty}\exp(ik_1\alpha_0nL)H_0^{(1)}(k_1\sqrt{(x'-x_n)^2+(z'-z)^2}). \tag{3.36}$$

The periodic Green's functions are also Floquet-periodic. For either Green's function, using (3.29) or (3.32) we have that

$$G^{(2p)}(\vec{x}',\vec{x}_n) = \exp(-ik_1\alpha_0 nL)G^{(2p)}(\vec{x}',\vec{x}), \tag{3.37}$$

where $\vec{x}_n = \vec{x} + \hat{i}nL$. Since the Floquet condition on any field function (3.8) has the conjugate phase of (3.37), the product of any Green's function times any field function ψ is periodic,

$$G^{(2p)}(\vec{x}',\vec{x}_n)\psi(\vec{x}_n) = G^{(2p)}(\vec{x}',\vec{x})\psi(\vec{x}). \tag{3.38}$$

This result will be used later to cancel vertical integrals in Green's theorem for the coordinate-space representation.

The three-dimensional Green's function in free space is given by

$$G^{(3)}(\vec{x}',\vec{x}) = \frac{1}{4\pi}\frac{\exp(ik_0|\vec{x}'-\vec{x}|)}{|\vec{x}'-\vec{x}|}, \tag{3.39}$$

where k_0 is a generic wave number. It satisfies the equation

$$(\nabla_3^2 + k_0^2)G^{(3)}(\vec{x}',\vec{x}) = -\delta(\vec{x}'-\vec{x}), \tag{3.40}$$

where $\vec{x} = (x,y,z)$. Its Fourier representation is

$$G^{(3)}(\vec{x}',\vec{x}) = \frac{1}{(2\pi)^3}\iiint \frac{\exp[ik_x(x'-x)+ik_y(y'-y)+ik_z(z'-z)]}{k^2-k_{0+}^2}dk_x dk_y dk_z, \tag{3.41}$$

with $k^2 = k_x^2+k_y^2+k_z^2$. Use complex integration on the preferred z-direction to yield

$$G^{(3)}(\vec{x}',\vec{x}) = \frac{i\pi}{(2\pi)^3}\iint \frac{\exp[ik_x(x'-x)+ik_y(y'-y)+iK_0|z'-z|]}{K_0}dk_x dk_y, \tag{3.42}$$

where

$$\begin{aligned} K_0 &= \sqrt{k_0^2-k_x^2-k_y^2}, \quad (k_x^2+k_y^2<k_0^2) && (3.43)\\ &= +i\sqrt{k_x^2+k_y^2-k_0^2}, \quad (k_x^2+k_y^2>k_0^2). && (3.44)\end{aligned}$$

In the upper region, let $k_0 = k_1$ in (3.42), and scale the wavenumbers as $k_x = k_1\alpha$ and $k_y = k_1\beta$. This yields the three-dimensional Green's function for region 1 in the Weyl representation

$$G_1^{(3)}(\vec{x}',\vec{x}) = \frac{i\pi k_1}{(2\pi)^3}\iint \frac{\exp[ik_1(\alpha(x'-x)+\beta(y'-y)+\gamma_1(\alpha,\beta)|z'-z|)]}{\gamma_1(\alpha,\beta)}d\alpha d\beta, \tag{3.45}$$

where

$$\begin{aligned} \gamma_1 &= \sqrt{1-\alpha^2-\beta^2}, \quad (\alpha^2+\beta^2<1), && (3.46)\\ &= +i\sqrt{\alpha^2+\beta^2-1}, \quad (\alpha^2+\beta^2>1). && (3.47)\end{aligned}$$

In the lower region, let $k_0 = k_2$ in (3.42), scale the wavenumbers the same (two-dimensional Snell's law) to yield the three-dimensional Green's function for region 2 in the Weyl representation

$$G_2^{(3)}(\vec{x}',\vec{x}) = \frac{i\pi k_1}{(2\pi)^3}\iint \frac{\exp[ik_1(\alpha(x'-x)+\beta(y'-y)+\gamma_2(\alpha,\beta)|z'-z|)]}{\gamma_2(\alpha,\beta)}d\alpha d\beta, \tag{3.48}$$

where

$$\gamma_2 = \sqrt{K^2-\alpha^2-\beta^2}, \quad (\alpha^2+\beta^2<K^2), \tag{3.49}$$

$$= +i\sqrt{\alpha^2+\beta^2-K^2}, \quad (\alpha^2+\beta^2>K^2). \tag{3.50}$$

The above results are for an infinite surface. To illustrate the reduction to a single cell of a two-dimensional periodic surface we choose a single layer potential (for either region) with density ψ which is any field function (here the normal derivative of the velocity potential)

$$(S\psi)(\vec{x}_h') = \iint_{-\infty}^{\infty} G^{(3)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dxdy. \tag{3.51}$$

Here the surface is doubly periodic (period L_1 in x and L_2 in y)

$$h(\vec{x}_\perp+\vec{x}_{n_1n_2}) = h(\vec{x}_\perp), \tag{3.52}$$

where n_1 and n_2 are integers and $\vec{x}_{n_1n_2} = \hat{i}n_1L_1 + \hat{j}n_2L_2$. The field function is Floquet-periodic in two dimensions, see (3.10). We can thus write (3.51) as a double sum over periodic cells

$$(S\psi)(\vec{x}_h') = \sum_{n_1=-\infty}^{\infty}\sum_{n_2=-\infty}^{\infty} I_{n_1n_2}(\vec{x}_h'), \tag{3.53}$$

where

$$I_{n_1n_2}(\vec{x}_h') = \int_{(2n_2-1)L_2/2}^{(2n_2+1)L_2/2}\int_{(2n_1-1)L_1/2}^{(2n_1+1)L_1/2} G^{(3)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dxdy. \tag{3.54}$$

Use the Weyl representation (3.45) or (3.48) for $G^{(3)}$, shift the integrations using $x''=x-n_1L_1$ and $y''=y-n_2L_2$ to yield

$$(S\psi)(\vec{x}_h') = \int_{-L_2/2}^{L_2/2}\int_{-L_1/2}^{L_1/2} G^{(3p)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h)dxdy, \tag{3.55}$$

where the three-dimensional periodic Green's function is given by

$$G^{(3p)}(\vec{x}_h',\vec{x}_h) = \frac{i\pi k_1}{(2\pi)^3}\iint \frac{\exp[ik_1(\alpha(x'-x)+\beta(y'-y)+\gamma|h(\vec{x}_\perp')-h(\vec{x}_\perp)|)]}{\gamma(\alpha,\beta)}P_1P_2d\alpha d\beta, \tag{3.56}$$

with the Poisson sums

$$P_1(\alpha) = \sum_{n_1=-\infty}^{\infty}\exp[in_1k_1L_1(\alpha_0-\alpha)] = \frac{2\pi}{k_1L_1}\sum_{j=-\infty}^{\infty}\delta(\alpha_j-\alpha), \tag{3.57}$$

and

$$P_2(\beta) = \sum_{n_2=-\infty}^{\infty}\exp[in_2k_1L_2(\beta_0-\beta)] = \frac{2\pi}{k_1L_2}\sum_{j'=-\infty}^{\infty}\delta(\beta_{j'}-\beta). \tag{3.58}$$

The grating equations are now $\alpha_j=\alpha_0+j\lambda/L_1$ and $\beta_{j'}=\beta_0+j'\lambda/L_2$. The result is the three-dimensional periodic Green's function for region 1 with both coordinates on the surface

$$G_1^{(3p)}(\vec{x}_h',\vec{x}_h) = \frac{i}{2k_1L_1L_2}\sum_{j=-\infty}^{\infty}\sum_{j'=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\beta_{j'}(y'-y)+\gamma_{1jj'}|h(\vec{x}_\perp')-h(\vec{x}_\perp)|)]}{\gamma_{1jj'}}, \tag{3.59}$$

where

$$\gamma_{1jj'} = \sqrt{1-\alpha_j^2-\beta_{j'}^2}, \quad (\alpha_j^2+\beta_{j'}^2<1) \tag{3.60}$$

$$= +i\sqrt{\alpha_j^2+\beta_{j'}^2-1}, \quad (\alpha_j^2+\beta_{j'}^2>1). \tag{3.61}$$

The three-dimensional periodic Green's function for region 2 is given by

$$G_2^{(3p)}(\vec{x}_h',\vec{x}_h) = \frac{i}{2k_1L_1L_2}\sum_{j=-\infty}^{\infty}\sum_{j'=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\beta_{j'}(y'-y)+\gamma_{2jj'}|h(\vec{x}_\perp')-h(\vec{x}_\perp)|)]}{\gamma_{2jj'}}, \tag{3.62}$$

where

$$\gamma_{2jj'} = \sqrt{K^2-\alpha_j^2-\beta_{j'}^2}, \quad (\alpha_j^2+\beta_{j'}^2<K^2) \tag{3.63}$$

$$= +i\sqrt{\alpha_j^2+\beta_{j'}^2-K^2}, \quad (\alpha_j^2+\beta_{j'}^2>K^2). \tag{3.64}$$

We use these Green's functions later for three-dimensional electromagnetic problems. Note again that the scaling is k_1 in front of both (3.59) and (3.62). Note also the obvious remark that for a two-dimensional surface $h(\vec{x}_\perp)$ we have two spectral parameters, j and j'.

Both Green's functions satisfy a two-dimensional Floquet condition

$$G^{(3p)}(\vec{x}_h',\vec{x}_h+\vec{x}_{n_1n_2}) = \exp[-ik_1(\alpha_0n_1L_1+\beta_0n_2L_2)]G^{(3p)}(\vec{x}_h',\vec{x}_h). \tag{3.65}$$

Combined with the two-dimensional Floquet condition on any field function (3.10), the product of any Green's function times a field function is periodic

$$G^{(3p)}(\vec{x}_h',\vec{x}_h+\vec{x}_{n_1n_2})\psi(\vec{x}_h+\vec{x}_{n_1n_2}) = G^{(3p)}(\vec{x}_h',\vec{x}_h)\psi(\vec{x}_h). \tag{3.66}$$

We use this property later to cancel side integrals in Green's theorem. Techniques for computing these periodic Green's functions are available[101,102].

3.4 Integral Methods in Coordinate Space for Scalar Problems

We first present the coordinate-space representation of the scattering from a periodic surface as comparison and contrast to that of the spectral representations in later sections. In addition, these yield rigorous representations for the scattered field above the highest surface excursion and for the transmitted field below the lowest surface excursion, as well as projections on lines above and below the surface.

The total field in region 1, ψ_1, equals the sum of incident plus scattered fields, $\psi_1 = \psi^{in}+\psi^{sc}$, and it satisfies the scalar Helmholtz equation

$$(\nabla_2^2+k_1^2)\psi_1(\vec{x}) = 0, \tag{3.67}$$

as do both incident and scattered fields. We do Green's theorem using ψ_1 and $G_1^{(2p)}$. Cross multiply (3.67) and (3.35), multiply by the characteristic function of region 1

$$\Theta_1(\vec{x}) = \theta(L/2-x)\theta(x+L/2)\theta(z-h(x))\theta(H_1-z), \tag{3.68}$$

where θ is the step function, $\theta(x) = 1$ when $x > 0$, and $\theta(x) = 0$ when $x < 0$, and integrate by parts. To express the results conveniently, introduce the bracket notation

$$[G_1^{(2p)}, \psi_1; \vec{x}', S] = \iint_S [G_1^{(2p)}(\vec{x}', \vec{x}_S)\partial_l \psi_1(\vec{x}_S) - \partial_l G_1^{(2p)}(\vec{x}', \vec{x}_S)] n_l ds, \tag{3.69}$$

where ∂_l is the partial derivative (∂_x for $l = 1$ and ∂_z for $l = 2$), n_l is the non-unit surface normal, and ds the arc length along the surface. Repeated subscripts are summed. There are four surfaces S: $x = \pm L/2$ ($h < z < H_1$), $z = h$, and $z = H_1$, both with $-L/2 < x < L/2$. The result is

$$\psi_1(\vec{x}')\Theta_1(\vec{x}') = [G_1^{(2p)}, \psi_1; \vec{x}', L/2] - [G_1^{(2p)}, \psi_1; \vec{x}', -L/2] + [G_1^{(2p)}, \psi_1; \vec{x}', H_1] - [G_1^{(2p)}, \psi_1; \vec{x}', h]. \tag{3.70}$$

Using (3.38), the first two brackets on the right hand side of (3.70) cancel by Floquet periodicity. For the moment assume the integral on H_1 represents the incident field (proof below), i.e.

$$\psi^{in}(\vec{x}') = [G_1^{(2p)}, \psi_1; \vec{x}', H_1]. \tag{3.71}$$

We thus have three results. Inside region 1, $h(x') < z' < H_1$, $\Theta_1 = 1$, we have

$$\psi_1(\vec{x}') = \psi^{in}(\vec{x}') - [G_1^{(2p)}, \psi_1; \vec{x}', h]. \tag{3.72}$$

Outside region 1, where $\Theta_1 = 0$, we have from (3.70)

$$\psi^{in}(\vec{x}') = [G_1^{(2p)}, \psi_1; \vec{x}', h], \tag{3.73}$$

which is an Extinction Theorem, and on the surface $z' = h(x')$, taking into account the discontinuity of the double layer potential in (3.72), we have

$$\frac{1}{2}\psi_1(\vec{x}'_h) = \psi^{in}(\vec{x}'_h) - [G_1^{(2p)}, \psi_1; \vec{x}'_h, h]. \tag{3.74}$$

In particular, we can write the scattered field above the highest surface excursion, $z' > max(h)$, using (3.72). The absolute value in the Green's function in (3.72) is not present, i.e. we use the representation

$$G_1^{(2p)}(\vec{x}', \vec{x}_h) = \frac{i}{2k_1 L} \sum_{j=-\infty}^{\infty} \frac{\exp[ik_1(\alpha_j(x' - x) + \gamma_{1j}(z' - h(x)))]}{\gamma_{1j}}. \tag{3.75}$$

The result is that the scattered field above the highest surface excursion can be written exactly as a plane wave expansion of upgoing waves

$$\psi^{sc}(\vec{x}') = \sum_{j=-\infty}^{\infty} A_j \exp[ik_1(\alpha_j x' + \gamma_{1j} z')], \tag{3.76}$$

where A_j can be written as the integral

$$A_j = \frac{1}{L}\int_{-L/2}^{L/2} A(j, x)\exp[-ik_1(\alpha_j x + \gamma_{1j} h(x))]dx, \tag{3.77}$$

and the integrand $A(j,x)$ is in terms of the boundary unknowns

$$A(j,x) = \frac{-i}{2k_1\gamma_{1j}}\left\{\frac{\partial\psi_1}{\partial n}(\vec{x}_h) + ik_1(\gamma_{1j} - \alpha_j h'(x))\psi_1(\vec{x}_h)\right\}. \tag{3.78}$$

We have written $A(j,x)$ as a function of two variables, the first a discrete spectral (S) variable j which has replaced the field coordinate variable, and the second a continuous coordinate (C) variable x, which is the surface integration variable. This is the basis for the spectral-coordinate (SC) approach used in Sec. 5.

There remains to prove (3.71). This time the representation for the Green's function evaluated on $z = H_1$ is with $\vec{x}_1 = (x, H_1)$

$$G_1^{(2p)}(\vec{x}', \vec{x}_1) = \frac{i}{2k_1 L}\sum_{j=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x) + \gamma_{1j}(H_1 - z'))]}{\gamma_{1j}}. \tag{3.79}$$

Using (3.79) in (3.71) and the representation (3.76) for the scattered field it is straighforward to show that

$$[G_1^{(2p)}, \psi^{sc}; \vec{x}', H_1] = 0. \tag{3.80}$$

Further, if we assume a general plane wave decomposition of the incident field in terms of downgoing waves

$$\psi^{in}(\vec{x}) = \sum_n I_n \exp[ik_1(\alpha_n x - \gamma_{1n} z)], \tag{3.81}$$

the relation

$$[G_1^{(2p)}, \psi^{in}; \vec{x}', H_1] = \psi^{in}(\vec{x}'), \tag{3.82}$$

follows immediately. Alternatively, one can view the integrand in (3.71)

$$\frac{\partial\psi_1}{\partial z}(x, H_1) - ik_1\gamma_{1j}\psi_1(x, H_1), \tag{3.83}$$

as projecting out only the downgoing waves and canceling the scattered waves. The combination of (3.80) and (3.82) is the proof of (3.71).

In the region below the surface, region 2, the total field ψ_2 satisfies the Helmholtz equation

$$(\nabla_2^2 + k_2^2)\psi_2(\vec{x}) = 0, \tag{3.84}$$

where the wavenumber $k_2 = Kk_1$ is written in terms of a scale factor K. The region is defined by the characteristic function

$$\Theta_2(\vec{x}) = \theta(L/2 - x)\theta(x + L/2)\theta(h(x) - z)\theta(z - H_2). \tag{3.85}$$

The Green's function $G_2^{(2p)}$ is given by

$$G_2^{(2p)}(\vec{x}', \vec{x}) = \frac{i}{2k_1 L}\sum_{j=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x) + \gamma_{2j}|z'-z|)]}{\gamma_{2j}}, \tag{3.86}$$

where γ_{2j} is defined in (3.33). Green's theorem in this region, the cancellation of integrals along $x = \pm L/2$ by Floquet periodicity, and the vanishing of the integral along H_2 since at this value of z the total field consists of downward propagating waves, yields the result

$$\psi_2(\vec{x}')\Theta_2(\vec{x}') = [G_2^{(2p)}, \psi_2; \vec{x}', h]. \tag{3.87}$$

On the boundary, the limit of (3.87) is

$$\frac{1}{2}\psi_2(\vec{x}'_h) = [G_2^{(2p)}, \psi_2; \vec{x}'_h, h]. \tag{3.88}$$

For $z' < min(h)$, (3.87) yields a representation of the total field in region 2 in terms of downward propagating plane waves

$$\psi_2(\vec{x}') = \sum_{j=-\infty}^{\infty} B_j \exp[ik_1(\alpha_j x' - \gamma_{2j} z')], \tag{3.89}$$

where

$$B_j = \frac{1}{L}\int_{-L/2}^{L/2} B(j,x)\exp[-ik_1(\alpha_j x - \gamma_{2j} h(x))]dx, \tag{3.90}$$

and $B(j,x)$ is in terms of the boundary values from region 2

$$B(j,x) = \frac{i}{2k_1\gamma_{2j}}\left\{\frac{\partial\psi_2(\vec{x}_h)}{\partial n} - ik_1(\gamma_{2j} + \alpha_j h'(x))\psi_2(\vec{x}_h)\right\}. \tag{3.91}$$

We have assumed that ψ is a velocity potential. Then the continuity conditions at the boundary are written as the continuity of velocity

$$\frac{\partial\psi_2}{\partial n}(\vec{x}_h) = \frac{\partial\psi_1}{\partial n}(\vec{x}_h) \doteq N(\vec{x}_h), \tag{3.92}$$

and continuity of pressure

$$\rho_2\psi_2(\vec{x}_h) = \rho_1\psi_1(\vec{x}_h), \tag{3.93}$$

where ρ_j are the densities. In (3.92) we defined the normal derivative boundary unknown as N, and we define the surface field boundary unknown as $\psi(\vec{x}_h) = \psi_1(\vec{x}_h)$, so that $\psi_2(\vec{x}_h) = \frac{1}{\rho}\psi(\vec{x}_h)$ where $\rho = \rho_2/\rho_1$. Using these unknowns, (3.78) and (3.91) become

$$A(j,x) = \frac{-i}{2k_1\gamma_{1j}}[N(\vec{x}_h) + ik_1(\gamma_{1j} - \alpha_j h'(x))\psi(\vec{x}_h)], \tag{3.94}$$

and

$$B(j,x) = \frac{i}{2k_1\gamma_{2j}}[N(\vec{x}_h) - \frac{ik_1}{\rho}(\gamma_{2j} + \alpha_j h'(x))\psi(\vec{x}_h)]. \tag{3.95}$$

We can summarize these results using single(S) and double(D) layer potentials

$$(S_j u)(\vec{x}') = \int_{-L/2}^{L/2} G_j^{(2p)}(\vec{x}', \vec{x}_h)u(\vec{x}_h)dx, \tag{3.96}$$

and

$$(D_j v)(\vec{x}') = \int_{-L/2}^{L/2} \frac{\partial G_j^{(2p)}}{\partial n}(\vec{x}', \vec{x}_h)v(\vec{x}_h)dx, \tag{3.97}$$

and write the integral equations (3.74) and (3.88) in symbolic form as

$$\frac{1}{2}\psi = \psi^{in} - (S_1 N) + (D_1\psi), \tag{3.98}$$

and

$$\frac{1}{2}\psi = \rho(S_2 N) - (D_2\psi). \tag{3.99}$$

Various combinations of these equations and integral equations formed by first taking the normal derivative of the field representations (3.72) and (3.87) and passing to the surface limit can be used to solve for the boundary unknowns ψ and N. For the Dirichlet problem, $\psi = 0$ and $\rho = 0$ so (3.99) disappears and (3.98) is an integral equation of first kind for N. For the Neumann problem, first divide (3.99) by ρ, then let $\rho \to \infty$ and set $N = 0$.

Direct integral equation methods have been used to computationally solve this problem[21,22,67]. Other integral equation solutions[36,37,38,39] have been compared to the solutions of spectral methods presented later in this paper. Other methods have also been employed[18,19,20,73,74,86]. Point collocation questions arise[10,25,53,60,72] for any coordinate based method.

3.5 Partial Spectral Methods for Scalar Problems

In this section we use a direct method to generate integral equations in a partial spectral representation. The method uses Green's theorem again, but not the Green's function. Define the up- and down-going plane wave states in region 1

$$\phi_{1j}^{\pm}(\vec{x}) = \exp[ik_1(-\alpha_j x \pm \gamma_{1j} z)], \tag{3.100}$$

which satisfy the same Helmholtz equation as ψ_1, (3.67),

$$(\nabla_2^2 + k_1^2)\phi_{1j}^{\pm}(\vec{x}) = 0. \tag{3.101}$$

For convenience, in (3.100) we have chosen the conjugate in the x-coordinate. In the Green's function this occurs naturally. Cross multiply (3.67) and (3.101) and subtract the results, multiply by Θ_1 from (3.68), integrate over all space, and then integrate by parts. Since all fields are Floquet periodic, the integrals along $x = \pm L/2$ cancel. The results can be expressed with the collapsed bracket notation

$$[u, v; S] = \int_S [u(\vec{x}_S)\partial_l v(\vec{x}_S) - v(\vec{x}_S)\partial_l u(\vec{x}_S)] n_l ds, \tag{3.102}$$

where, unlike the bracket notation in Sec.4, no exterior coordinate-space variable appears. There are two surfaces, $z = h$ with $-L/2 < x < L/2$, and $z = H_1$ with $-L/2 < x < L/2$. The result is

$$[\phi_{1j}^{\pm}, \psi_1; h] = [\phi_{1j}^{\pm}, \psi_1; H_1]. \tag{3.103}$$

The result can be thought of as an analytic continuation from the periodic surface $z = h$ to a flat plane $z = H_1$ above the surface. The right hand side of (3.103) can be evaluated explicitly using (3.76) and (3.81) to give

$$[\phi_{1j}^{\pm}, \psi_1; H_1] = 2ik_1 L\gamma_{1j}\left\{\begin{matrix} -I_j \\ A_j \end{matrix}\right\}. \tag{3.104}$$

Here the up-going plane waves ϕ_{1j}^{+} project out the down-going spectral components of the incident wave I_j, and the down-going plane waves ϕ_{1j}^{-} project out the up-going spectral components of the scattered waves A_j. Combining this with the left hand side of (3.103) we get the set of equations for region 1

$$\frac{1}{L}\int_{-L/2}^{L/2} \phi_{1j}^{\pm}(\vec{x}_h) U_j^{\pm}(\vec{x}_h) dx = \gamma_{1j}\left\{\begin{matrix} -I_j \\ A_j \end{matrix}\right\}, \tag{3.105}$$

where

$$U_j^{\pm}(\vec{x}_h) = \frac{1}{2ik_1}[N(\vec{x}_h) - ik_1(\pm\gamma_{1j} + \alpha_j h'(x))\psi(\vec{x}_h)]. \tag{3.106}$$

We have incorporated the boundary unknowns defined following (3.92). Note that $U_j^- = \gamma_{1j}A(j,x)$ from (3.94).

In region 2, the up- and down-going plane waves are given by

$$\phi_{2j}^{\pm}(\vec{x}) = \exp[ik_1(-\alpha_j x \pm \gamma_{2j} z)], \tag{3.107}$$

which satisfy the Helmholtz equation

$$(\nabla_2^2 + k_2^2)\phi_{2j}^{\pm}(\vec{x}) = 0. \tag{3.108}$$

Cross multiply (3.84) and (3.108), multiply the result by Θ_2 from (3.85), integrate over all space, then integrate by parts. The Floquet periodicity cancels the integrals on $x = \pm L/2$ and the result is the analytic continuation

$$[\phi_{2j}^{\pm}, \psi_2; h] = [\phi_{2j}^{\pm}, \psi_2; H_2]. \tag{3.109}$$

The right hand side of (3.109) can be evaluated using (3.89) for ψ_2 to yield

$$[\phi_{2j}^{\pm}, \psi_2; H_2] = -2ik_1 L\gamma_{2j}\left\{{B_j \atop 0}\right\}. \tag{3.110}$$

Combined with (3.109), and incorporating the definitions of the boundary values following (3.92) yields the equations from the lower region

$$\frac{1}{L}\int_{-L/2}^{L/2} \phi_{2j}^{\pm}(\vec{x}_h) L_j^{\pm}(\vec{x}_h) dx = -\gamma_{2j}\left\{{B_j \atop 0}\right\}. \tag{3.111}$$

where

$$L_j^{\pm}(\vec{x}_h) = \frac{1}{2ik_1}[N(\vec{x}_h) - \frac{ik_1}{\rho}(\pm\gamma_{2j} + \alpha_j h'(x))\psi(\vec{x}_h)]. \tag{3.112}$$

The lower equation in (3.111) is a spectral version of the Extinction Theorem.

The procedure is to solve the combined U^+ equation in (3.105) and the L^- equation in (3.111) for the boundary unknowns N and ψ, and to evaluate the U^- and L^+ equations for the scattered (A_j) and transmitted (B_j) amplitudes. The scattered and transmitted fields can be then found from (3.76) and (3.89) respectively. In order to find field values in the surface wells, we must use these boundary unknowns in (3.72) and (3.87) respectively.

The advantage of the method is that there are no Green's functions to compute. Instead, the results are projected onto plane wave based basis functions (really Rayleigh functions since they're on the surface). The Green's function does this in an alternate way.

It is useful with any theory to check simple special cases. It is also necessary that the general results reduce to simple solvable cases. Here we take the flat surface limit ($h = 0$), and derive from them the Fresnel reflection and transmission coefficients as a necessary check on the general results. For $h = 0$, let $L \to \infty$, so that for any finite j, $\lim_{L\to\infty} \alpha_j = \alpha_0$, so that the only surviving waves are the 0^{th} order reflection (A_0) and transmission (B_0) amplitudes. The surface fields N and ψ thus have two different flat-surface field representations which are

$$\psi_1(x,0) = (I_0 + A_0)\exp[ik_1\alpha_0 x], \tag{3.113}$$

$$N_1(x,0) = -ik_1\gamma_{10}(I_0 - A_0)\exp[ik_1\alpha_0 x], \tag{3.114}$$

$$\psi_2(x,0) = B_0\exp[ik_1\alpha_0 x], \tag{3.115}$$

and

$$N_2(x,0) = -ik_1\gamma_{20}B_0\exp[ik_1\alpha_0 x]. \tag{3.116}$$

From (3.105) and (3.111) we have

$$A_0 = \frac{1}{2ik_1\gamma_{10}}\lim_{L\to\infty}\frac{1}{L}\int_{-L/2}^{L/2}[N(x,0) + ik_1\gamma_{10}\psi(x,0)]\exp[-ik_1\alpha_0 x]dx, \tag{3.117}$$

and

$$B_0 = \frac{-1}{2ik_1\gamma_{20}}\lim_{L\to\infty}\frac{1}{L}\int_{-L/2}^{L/2}[N(x,0) - \frac{ik_1}{\rho}\gamma_{20}\psi(x,0)]\exp[-ik_1\alpha_0 x]dx. \tag{3.118}$$

If we use the flat-surface field representations on the surface from region 1, (N_1 and ψ_1), in the A_0 equation, and the flat-surface field representations from region 2, (N_2 and ψ_2), in the B_0 equation, we just get identities. Instead, use the opposite procedure, i.e. write A_0 and B_0 as

$$A_0 = \frac{1}{2ik_1\gamma_{10}}\lim_{L\to\infty}\frac{1}{L}\int_{-L/2}^{L/2}[N_2(x,0) + ik_1\gamma_{10}\psi_2(x,0)]\exp[-ik_1\alpha_0 x]dx, \tag{3.119}$$

and

$$B_0 = \frac{-1}{2ik_1\gamma_{20}}\lim_{L\to\infty}\frac{1}{L}\int_{-L/2}^{L/2}[N_1(x,0) - \frac{ik_1}{\rho}\gamma_{20}\psi_1(x,0)]\exp[-ik_1\alpha_0 x]dx. \tag{3.120}$$

Using (3.113) through (3.116) in (3.119) and (3.120) we get two equations

$$A_0 = \frac{\rho\gamma_{10} - \gamma_{20}}{2\gamma_{10}}B_0, \tag{3.121}$$

and

$$B_0 = \frac{\gamma_{10}[I_0 - A_0] + (\gamma_{20}/\rho)[I_0 + A_0]}{2\gamma_{10}}. \tag{3.122}$$

These can be solved to yield the Fresnel reflection coefficient

$$\frac{A_0}{I_0} = \frac{\rho\gamma_{10} - \gamma_{20}}{\rho\gamma_{10} + \gamma_{20}}, \tag{3.123}$$

and the Fresnel transmission coefficient

$$\frac{B_0}{I_0} = \frac{2\gamma_{10}}{\rho\gamma_{10} + \gamma_{20}}. \tag{3.124}$$

Finally we have that

$$1 + \frac{A_0}{I_0} = \rho\frac{B_0}{I_0}, \tag{3.125}$$

as expected.

3.6 Surface Inversion Using the Partial Spectral Method

We can use the partial spectral results from Sec.5 to develop simple algorithms to reconstruct the surface height $h(x)$ from the knowledge of the incident and scattered field amplitudes I_j and A_j. For simplicity, we choose the Dirichlet problem, $\psi(\vec{x}_h) = 0$. This is a perfectly reflecting case, and (3.111) vanishes identically (multiply it by ρ, and then set $\rho = 0$). The resulting equations (3.105) become

$$\frac{1}{L}\int_{-L/2}^{L/2} \phi_{1j}^{\pm}(\vec{x}_h) N(\vec{x}_h) dx = 2ik_1\gamma_{1j}\left\{\begin{smallmatrix} -I_j \\ A_j \end{smallmatrix}\right\}. \tag{3.126}$$

We describe two methods, the first is perturbation theory in the surface height, and the second is the use of the Kirchhoff approximation for the normal derivative N. The full details of both methods with numerical results were presented in[111,112]. There the methods were applied to truncated rough surfaces. Some other methods can be found in[59] for uniqueness questions and[17] for more detailed reconstruction algorithms.

For perturbation theory (3.100) is used on the surface, and becomes

$$\phi_{1j}^{\pm}(\vec{x}_h) \approx \exp[-ik_1\alpha_j x](1 \pm ik_1\gamma_{1j}h(x)). \tag{3.127}$$

Substituting (3.127) in (3.126), and adding and subtracting the resulting equations yields the two results

$$\frac{1}{L}\int_{-L/2}^{L/2} \exp[-ik_1\alpha_j x] N(\vec{x}_h) dx = -ik_1\gamma_{1j}(I_j - A_j), \tag{3.128}$$

and

$$\frac{1}{L}\int_{-L/2}^{L/2} \exp[-ik_1\alpha_j x] N(\vec{x}_h) h(x) dx = -(I_j + A_j). \tag{3.129}$$

Fourier inverting both equations yields

$$N(\vec{x}_h) = -ik_1 \sum_{j=-\infty}^{\infty} \exp[ik_1\alpha_j x](I_j - A_j), \tag{3.130}$$

and

$$N(\vec{x}_h)h(x) = -\sum_{j=-\infty}^{\infty} \exp[ik_1\alpha_j x](I_j + A_j). \tag{3.131}$$

Divide (3.131) by (3.130) (so that we factor out the boundary condition) and take the real part to get the approximation to the surface profile h_{PT} produced by perturbation theory

$$h_{PT}(x) = \frac{1}{k_1} Im\left\{ \frac{\sum_{j=-\infty}^{\infty}(I_j + A_j]\exp[ik_1\alpha_j x]}{\sum_{j=-\infty}^{\infty}(I_j - A_j)\exp[ik_1\alpha_j x]} \right\}. \tag{3.132}$$

where (Im) is the imaginary part. The equation simplifies for a single incident wave ($I_j = \delta_{j0}I_0$) to be

$$h_{PT}(x) = \frac{1}{k_1} Im\left\{ \frac{I_0 + \sum_{j=-\infty}^{\infty} A_j \exp[i2\pi jx/L]}{I_0 - \sum_{j=-\infty}^{\infty} A_j \exp[i2\pi jx/L]} \right\}. \tag{3.133}$$

These equations (3.132) and (3.133) express the surface in terms of the amplitudes of the incident and scattered fields.

For the Kirchhoff approximation (KA), assume a single plane wave incidence

$$\psi^{in}(\vec{x}) = I_0 \exp[ik_1(\alpha_0 x - \gamma_{10} z)], \tag{3.134}$$

and approximate the normal derivative on the surface in (3.126) by twice the normal derivative of the incident field

$$N(\vec{x}_h) \approx N^{KA}(\vec{x}_h) = 2n_l \partial_l \psi^{in}(\vec{x}_h). \tag{3.135}$$

For the lower equation in (3.126) this yields

$$\frac{1}{L}\int_{-L/2}^{L/2} [\gamma_{10} + \alpha_0 h'(x)] \exp[-ik_1(p_j x + q_j h(x))] dx = -\gamma_{1j} A_j / I_0, \tag{3.136}$$

where

$$p_j = \alpha_j - \alpha_0, \tag{3.137}$$

and

$$q_j = \gamma_{1j} + \gamma_{0j}. \tag{3.138}$$

The $h'(x)$ term in (3.136) can be integrated by parts to yield

$$\frac{1}{L}\int_{-L/2}^{L/2} \exp[-ik_1(p_j x + q_j h(x))] dx = -f_j^- A_j / I_0, \tag{3.139}$$

where

$$f_j^- = \frac{\gamma_{1j}(\gamma_{1j} + \gamma_{10})}{\gamma_{1j}\gamma_{10} + (1 - \alpha_j \alpha_0)}. \tag{3.140}$$

We can re-express p_j and q_j using trig identities as

$$p_j = \sin(\theta_j^{sc}) - \sin(\theta^{in}) = 2\cos\left\{\frac{\theta_j^{sc} + \theta^{in}}{2}\right\} \sin\left\{\frac{\theta_j^{sc} - \theta^{in}}{2}\right\}, \tag{3.141}$$

and

$$q_j = \cos(\theta_j^{sc}) + \cos(\theta^{in}) = 2\cos\left\{\frac{\theta_j^{sc} + \theta^{in}}{2}\right\} \cos\left\{\frac{\theta_j^{sc} - \theta^{in}}{2}\right\}. \tag{3.142}$$

We thus have that p_j and q_j are confined to an Ewald circle

$$p_j^2 + q_j^2 \leqslant 4. \tag{3.143}$$

and we further have $|p_j| \leqslant 2$ and $|q_j| \leqslant 2$. This restricts the acceptable j values to a set J and correspondingly restricts the acceptable scattering angles (modulo the incident angle), and thus the scattered amplitudes and fields used for the inversion. For fixed q_j, say q_{j_1}, (3.139) is a periodic Fourier transform restricted in p_j and thus restricted in the set J. As q_{j_1} increases, p_j decreases, which is equivalent to a low-pass filter. As q_{j_1} decreases, more data near grazing illumination and scattering is involved, where the Kirchhoff approximation gets worse. For fixed q_{j_1}, assume the integral in (3.139) can be approximately inverted to yield

$$\exp[-ik_1 q_{j_1} h(x)] = \frac{-1}{I_0} \sum_J f_j^- A_j \exp[ik_1 p_j x] \doteq \mathscr{R}(x). \tag{3.144}$$

Taking the real *Re* and imaginary *Im* parts of (3.144) (and neglecting periodic phase shifts) yields h_{KA}, the Kirchhoff approximation of the surface height

$$h_{KA}(x,q_{j_1}) = \frac{1}{k_1 q_{j_1}} \arctan\left\{\frac{-Im(\mathscr{R}(x))}{Re(\mathscr{R}(x))}\right\}, \tag{3.145}$$

which again produces the surface height function in terms of the scattered field amplitudes this time modulated by the Kirchhoff components. Each q_{j_1} produces a different value of h_{KA}. For a non-periodic truncated random surface the method was used successfully to reconstruct ensemble surface height functions with approximately twice the rms height as for perturbation theory[112]. The cited paper also contains a discussion of the various angle combinations for different reconstructions.

3.7 Full Spectral Methods for Scalar Problems: Physical Optics Modified Fourier Basis and Floquet-Fourier Expansions

In Sec.4, both the exterior and interior (integration) variables were in coordinate space. The equations generated were formally exact for the solution of the boundary values $\psi(\vec{x}_h)$ and $N(\vec{x}_h)$. Once the boundary values were found, the scattered and transmitted fields anywhere away from the surface wells could be evaluated via either direct transforms or summation methods in the resulting plane wave cum evanescent wave expansions. The periodic Green's function was used and had to be computed. Acceleration methods to do this are available[101,102].

In Sec.5, we used plane/evanescent waves to derive another set of equations, again formally exact, for the boundary unknowns which avoided the use of the periodic Green's functions. The equations to be solved were similar to the equations to be evaluated in the sense that both involved a close interplay between spectral and coordinate parameters in "parallel" as distinct from the "serial" presentation of methods in Sec.4.

Solution of the boundary unknowns in Secs.4 and 5 using direct discretization methods involves matrix inversion where the rows and columns of the matrix are both sampled in coordinate space, and the sampling methods are flexible. In Sec.5 the columns are sampled in coordinate space, but the rows are sampled in spectral space, and this is proscribed in terms of the Bragg waves. Convergence and the usefulness of the two sets of solutions have been discussed[36,37,38,39]. The major point is that the limits of convergence, stability and errors are numerical and directly related to the solution of *exact formal equations* and not to any strictly "physical" approximations.

That changes when we attempt to approximate the surface fields in some spectral basis, and thus to write equations fully in spectral space. The first question is what do we mean by spectral space in this context? The second is what do we know about possible expansions? The main thing we know is that the surface fields are the limits of Floquet-periodic functions, so they must also be Floquet-periodic. In particular, they should not be expanded in a pure Fourier series (no matter the temptation) since the latter are only valid for normal incidence ($\alpha_0 = 0$), where the Floquet periodicity reduces to ordinary periodicity. The validity of a pure Fourier expansion deteriorates for non-normal incidence.

In this section we briefly describe the use of a pure Floquet type expansion which defines "spectrum" in one particular way. It is also a physical optics (PO) expansion explained below. From this we are able to infer the results for what we refer to as a Floquet-Fourier (FF) expansion, and these latter results are presented at the end of the section. The PO expansions for the

boundary unknowns are

$$\psi(\vec{x}_h) = \exp[-ik_1\gamma_{10}h(x)] \sum_{j'=-\infty}^{\infty} \psi_{j'}^{(PO)} \exp[ik_1\alpha_{j'}x], \tag{3.146}$$

or, written in another form

$$\psi(\vec{x}_h) = \exp[ik_1\alpha_0 x - ik_1\gamma_{10}h(x)] \sum_{j'=-\infty}^{\infty} \psi_{j'}^{(PO)} \exp[i2\pi j'x/L], \tag{3.147}$$

where the term outside the summation can be written using the complex conjugate of (3.100)

$$\exp[ik_1\alpha_0 x - ik_1\gamma_{10}h(x)] = \bar{\phi}_{10}^{+}(\vec{x}_h). \tag{3.148}$$

The term is the physical optics or Kirchhoff approximation of a down-going plane wave evaluated on the boundary. It serves to modulate the remaining Fourier series, and can be viewed as a precursor to more general Waterman-type expansions[105] in terms of down-going waves. (If the h term is not present in (3.146), the expansion is still a Floquet-periodic expansion, and, since it is a generalization of the Fourier expansion, has the advantage of being invertible. Its result can be inferred from the results below, and are presented at the end of this section.) The normal derivative is similarly expanded

$$N(\vec{x}_h) = ik_1 \exp[-ik_1\gamma_{10}h(x)] \sum_{j'=-\infty}^{\infty} N_{j'}^{(PO)} \exp[ik_1\alpha_{j'}x]. \tag{3.149}$$

Here we have scaled the normal derivative term by ik_1 for convenience. The expansion was initially introduced as a physical optics modified Fourier expansion[31,32,33], and used by several others[15,26,27,55,106,107]. The reference[33] can be viewed as the exact version of the approximate Rayleigh-Fano equations[97] valid in perturbation theory for shallow surfaces. The expansions can be substituted into (3.105) and (3.111) to yield

$$\sum_{j'=-\infty}^{\infty} M_{1jj'}^{\pm}(PO)[N_{j'}^{(PO)} \mp \gamma_{1j}\psi_{j'}^{(PO)}] - \alpha_j \sum_{j'=-\infty}^{\infty} \tilde{M}_{1jj'}^{\pm}(PO)\psi_{j'}^{(PO)} = 2\gamma_{1j}\left\{{}^{-I_j}_{A_j}\right\}, \tag{3.150}$$

and

$$\sum_{j'=-\infty}^{\infty} M_{2jj'}^{\pm}(PO)[N_{j'}^{(PO)} \mp \frac{ik_1}{\rho}\gamma_{2j}\psi_{j'}^{(PO)}] - \frac{ik_1}{\rho}\alpha_j \sum_{j'=-\infty}^{\infty} \tilde{M}_{2jj'}^{\pm}(PO)\psi_{j'}^{(PO)} = -2\gamma_{2j}\left\{{}^{B_j}_{0}\right\}, \tag{3.151}$$

where the physical optics (PO) matrix elements are $(p = 1,2)$

$$M_{pjj'}^{\pm}(PO) = \frac{1}{L}\int_{-L/2}^{L/2} \exp(ik_1[(\pm\gamma_{pj} - \gamma_{10})h(x) + (\alpha_{j'} - \alpha_j)x])dx, \tag{3.152}$$

(note that $M_{10j'}^{+}(PO) = \delta_{j'0}$) and

$$\tilde{M}_{pjj'}^{\pm}(PO) = \frac{1}{L}\int_{-L/2}^{L/2} h'(x)\exp(ik_1[(\pm\gamma_{pj} - \gamma_{10})h(x) + (\alpha_{j'} - \alpha_j)x])dx. \tag{3.153}$$

The latter is written in such a way that integration by parts is obvious. Using integration by parts, the equations reduce to a simple form

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{1jj'}(PO)[N^{(PO)}_{j'} \mp a^{\pm}_{1jj'}\psi^{(PO)}_{j'}] = 2\gamma_{1j}\{{}^{-I_j}_{A_j}\}, \tag{3.154}$$

where

$$a^{\pm}_{1jj'} = \frac{\pm(1-\alpha_j\alpha_{j'}) - \gamma_{1j}\gamma_{10}}{\pm\gamma_{1j} - \gamma_{10}}, \tag{3.155}$$

and

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{2jj'}(PO)[N^{(PO)}_{j'} \mp a^{\pm}_{2jj'}\psi^{(PO)}_{j'}] = -2\gamma_{2j}\{{}^{B_j}_{0}\}, \tag{3.156}$$

where

$$a^{\pm}_{2jj'} = \frac{\pm(K^2-\alpha_j\alpha_{j'}) - \gamma_{2j}\gamma_{10}}{\rho(\pm\gamma_{2j} - \gamma_{10})}. \tag{3.157}$$

Finally, it is useful to rewrite the physical optics matrix elements as

$$M^{\pm}_{pjj'}(PO) = \frac{1}{L}\int_{-L/2}^{L/2} \exp[-i2\pi(j-j')x/L + ik_1(\pm\gamma_{pj} - \gamma_{10})h(x)]dx, \tag{3.158}$$

which displays the Fourier part explicitly. Note that for a flat surface, the only elements of (3.158) which survive are the diagonal elements $j = j'$ (which equal 1). Further, $\tilde{M}^{\pm}_{pjj'}(PO) = 0$, $a^{\pm}_{1jj'} = \gamma_{1j}$, $a^{\pm}_{2jj'} = \gamma_{2j}/\rho$, and, for a single plane wave incidence ($I_j = \delta_{j0}$), the usual flat surface limit of (3.154) and (3.156) follows directly.

A Floquet-Fourier (FF) expansion for the boundary unknowns can be written as

$$\psi(\vec{x}_h) = \sum_{j'=-\infty}^{\infty} \psi^{(FF)}_{j'} \exp(ik_1\alpha_{j'}x), \tag{3.159}$$

and

$$N(\vec{x}_h) = ik_1 \sum_{j'=-\infty}^{\infty} N^{(FF)}_{j'} \exp(ik_1\alpha_{j'}x). \tag{3.160}$$

The equations corresponding to (3.154) and (3.156) are thus

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{1jj'}(FF)[N^{(FF)}_{j'} \mp a_{1jj'}\psi j'^{(FF)}] = 2\gamma_{1j}\{{}^{-I_j}_{A_j}\}, \tag{3.161}$$

and

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{2jj'}(FF)[N^{(FF)}_{j'} \mp a_{2jj'}\psi^{(FF)}_{j'}] = -2\gamma_{2j}\{{}^{B_j}_{0}\}, \tag{3.162}$$

where

$$a_{1jj'} = \frac{1-\alpha_j\alpha_{j'}}{\gamma_{1j}}, \tag{3.163}$$

$$a_{2jj'} = \frac{K^2-\alpha j\alpha_{j'}}{\rho\gamma_{2j}}, \tag{3.164}$$

and, for $p = 1,2$, the matrix elements are

$$M^{\pm}_{pjj'}(FF) = \frac{1}{L}\int_{-L/2}^{L/2} \exp[-i2\pi(j-j')x/L \pm ik_1\gamma_{pj}h(x)]dx. \tag{3.165}$$

The FF equations follow from (3.154) through (3.158) by setting the γ_{10} term to zero. These provide an alternative set of equations to solve for the alternative boundary function coefficients to produce the same coefficients for the scattered and transmitted fields[58].

3.8 Full Spectral Methods for Scalar Problems: Conjugate Rayleigh Basis

A further spectral expansion consists in modifying the physical optics expansion by making the single physical optics plane wave dependent on the Bragg mode, so that the phase height term is dependent on the mode, and this leads to a conjugate Rayleigh (CR) expansion using the complex conjugate of the plane wave states (3.100) evaluated on the surface as

$$\psi(\vec{x}_h) = \sum_{j'=-\infty}^{\infty} \psi^{(CR)}_{j'} \exp[ik_1\alpha_{j'}x - ik_1\bar{\gamma}_{1j'}h(x)] = \sum_{j'=-\infty}^{\infty} \psi^{(CR)}_{j'}\bar{\phi}^{+}_{1j'}(\vec{x}_h), \tag{3.166}$$

and the scaled expansion for the normal derivative

$$N(\vec{x}_h) = ik_1\sum_{j'=-\infty}^{\infty} N^{(CR)}_{j'}\bar{\phi}^{+}_{1j'}(\vec{x}_h), \tag{3.167}$$

where the overbar is complex conjugation. Substituting these expansions in (3.105) and (3.111), and carrying out the integration by parts necessary to simplify the slope terms as in Sec.7 yields equations similar in form to (3.154) and (3.156). For the upper region equation we get

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{1jj'}(CR)[N^{(CR)}_{j'} \mp b^{\pm}_{1jj'}\psi^{(CR)}_{j'}] = 2\gamma_{1j}\left\{{-I_j \atop A_j}\right\}, \tag{3.168}$$

where

$$b^{\pm}_{1jj'} = \frac{1-\alpha_j\alpha_{j'} \mp \gamma_{1j}\bar{\gamma}_{1j'}}{\gamma_{1j} \mp \bar{\gamma}_{1j'}}, \tag{3.169}$$

and the matrix elements are defined as

$$M^{\pm}_{1jj'}(CR) = \frac{1}{L}\int_{-L/2}^{L/2} \exp[-i2\pi(j-j')x/L + ik_1(\pm\gamma_{1j} - \bar{\gamma}_{1j'})h(x)]dx = <\phi^{\pm}_{1j}, \phi^{+}_{1j'}>. \tag{3.170}$$

It is obvious that $M^{+}_{1jj'}(CR)$ is self-adjoint, positive definite and thus invertible, and this fact was used with success in solving the Dirichlet problem[3]. That is,

$$[M^{+}_{1jj'}]^{\star}(CR) = M^{+}_{1jj'}(CR), \tag{3.171}$$

where the symbol $\star$ represents the adjoint.

The same expansion for the equations in region 2 yields the equations

$$\sum_{j'=-\infty}^{\infty} M^{\pm}_{2jj'}(CR)[N^{(CR)}_{j'} \mp b^{\pm}_{2jj'}\psi^{(CR)}_{j'}] = -2\gamma_{2j}\left\{{B_j \atop 0}\right\}, \tag{3.172}$$

where

$$b^{\pm}_{2jj'} = \frac{1}{\rho}\frac{K_2 - \alpha_j\alpha_{j'} \mp \gamma_{2j}\bar{\gamma}_{1j'}}{\gamma_{2j} \mp \bar{\gamma}_{1j'}}, \tag{3.173}$$

and the matrix elements are

$$M^{\pm}_{2jj'}(CR) = \frac{1}{L}\int_{-L/2}^{L/2} \exp[-i2\pi(j-j')x/L + ik_1(\pm\gamma_{2j} - \bar{\gamma}_{1j'})h(x)]dx = <\phi^{\pm}_{2j}, \phi^{+}_{1j'}> . \tag{3.174}$$

3.9 Integral Equation Methods in Coordinate Space for Electromagnetic Problems

Up to now we have considered one-dimensional surfaces and acoustic problems. These correspond directly to electromagnetic scattering problems where there is no change in polarization for the scattered and transmitted fields. The general electromagnetic problem for a periodic dielectric interface is for a two-dimensional surface $z = h(\vec{x}_\perp) = h(x,y)$ which separates media of different dielectric constants ε_j for $j = 1,2$ and permeability μ_j. The wave numbers for the two regions are $k_j = k_0\sqrt{\varepsilon_j\mu_j}$ where $k_0 = \omega/c$, ω is the circular frequency and c the speed of light. There is now a change in polarization in the scattered and transmitted fields.

For the source-free electric field $\partial_i E_i = 0$, and each component of the electric field E_i with $i = 1,2,3$ satisfies the same Helmholtz equation as the scalar field, viz. in region 1

$$(\nabla_3^2 + k_1^2)E_{1i}(\vec{x}) = 0, \tag{3.175}$$

where E_{1i} is the i^{th} electric field component in region 1, and the Laplacian is three-dimensional. This is just the vector analogue of (3.67). We can use this to write the vector analogues of the scalar equations in Sec.4, using Green's theorem, the three-dimensional periodic Green's functions $G^{(3p)}$ from (3.59) and (3.62), the two-dimensional Floquet periodicity of the field, and the characteristic function defining the region, e.g. for region 1

$$\Theta_1(\vec{x}) = \theta(L_1/2 - x)\theta(x + L_1/2)\theta(L_2/2 - y)\theta(y + L_2/2)\theta(z - h(\vec{x}_\perp))\theta(H_1 - z). \tag{3.176}$$

In region 1 the result is

$$E_{1i}(\vec{x}')\Theta_1(\vec{x}') = E_i^{in}(\vec{x}') - [G_1^{(3p)}, E_{1i}; \vec{x}', h], \tag{3.177}$$

where E_i^{in} is the incident field, and the two-dimensional bracket is explicitly

$$[G_1^{(3p)}, E_{1i}; \vec{x}', h] = \iint_D [G_1^{(3p)}(\vec{x}', \vec{x}_h)N_{1i}(\vec{x}_h) - N_1^{(3p)}(\vec{x}', \vec{x}_h)E_{1i}(\vec{x}_h)]d\vec{x}_\perp, \tag{3.178}$$

where $\vec{x}_\perp = (x,y)$, the domain of integration D is $x\varepsilon[-L_1/2, L_1/2]$, $y\varepsilon[-L_2/2, L_2/2]$, and the normal derivatives are

$$N_{1i}(\vec{x}_h) = n_l(\vec{x}_\perp)\partial_l E_{1i}(\vec{x}_h), \tag{3.179}$$

and

$$N_1^{(3p)}(\vec{x}', \vec{x}_h) = n_l(\vec{x}_\perp)\partial_l G_1^{(3p)}(\vec{x}', \vec{x}_h), \tag{3.180}$$

the normal derivatives of the boundary unknown and the periodic Green's function respectively. From (3.177), the field representation in D is found by setting $\Theta_1 = 1$, the Extinction Theorem by setting $\Theta_1 = 0$, the scattered field is just the bracket term

$$E_{1i}^{sc}(\vec{x}') = -[G_1^{(3p)}, E_{1i}; \vec{x}', h], \tag{3.181}$$

and the boundary integral equation is

$$\frac{1}{2}E_{1i}(\vec{x}_h') = E_i^{in}(\vec{x}_h') - \iint_D [G_1^{(3p)}(\vec{x}_h', \vec{x}_h) N_{1i}(\vec{x}_h) - N_1^{(3p)}(\vec{x}_h', \vec{x}_h) E_{1i}(\vec{x}_h)] d\vec{x}_\perp, \tag{3.182}$$

with the boundary unknowns E_{1i} and its normal derivative N_{1i}.

Green's theorem in region 2 yields the representation for the total transmitted field

$$E_{2i}(\vec{x}')\Theta_2(\vec{x}') = [G_2^{(3p)}, E_{2i}; \vec{x}', h], \tag{3.183}$$

where $\Theta_2 = 1 - \Theta_1$, and the boundary integral equation becomes

$$\frac{1}{2}E_{2i}(\vec{x}_h') = \iint_D [G_2^{(3p)}(\vec{x}_h', \vec{x}_h) N_{2i}(\vec{x}_h) - N_2^{(3p)}(\vec{x}_h', \vec{x}_h) E_{2i}(\vec{x}_h)] d\vec{x}_\perp, \tag{3.184}$$

with boundary unknowns E_{2i} and its normal derivative N_{2i}. Equations (3.182) and (3.184) are similar to the scalar equations (3.74) and (3.88), but the fields and their normal derivatives are not the usual electromagnetic boundary values, the latter being typically written in terms of normal field components and currents[61]. So to continue we must relate these usual boundary conditions to our boundary unknowns.

For the electric field on the boundary we have the continuity condition of the normal component of the displacement vector $\vec{D} = \varepsilon \vec{E}$ which becomes

$$\varepsilon \vec{n} \cdot \vec{E}_2 = \vec{n} \cdot \vec{E}_1, \tag{3.185}$$

where $\varepsilon = \varepsilon_2/\varepsilon_1$, and the continuity of the magnetic current

$$\vec{n} \times \vec{E}_2 = \vec{n} \times \vec{E}_1. \tag{3.186}$$

These are four equations, three of which are independent. These three can be solved directly, or the four equations solved using a Moore-Penrose pseudo inverse to yield the boundary conditions on the electric field (in index notation) as

$$E_{2i}(\vec{x}_h) = C_{ij}(\vec{x}_h) E_{1j}(\vec{x}_h), \tag{3.187}$$

with repeated subscripts summed from 1 to 3 and

$$C_{ij}(\vec{x}_h) = \delta_{ij} + (\varepsilon^{-1} - 1)\hat{n}_i \hat{n}_j, \tag{3.188}$$

with $\hat{n}$ representing the unit normal. These boundary conditions were introduced some time ago[35] and used successfully for scattering from a body of revolution[41].

The continuity conditions on the normal derivative components are more involved. The full details are in[35]. Briefly we introduce the bracket notation for when we set the field on the surface first and then differentiate

$$\{E_m\} \doteq E_m(x, y, h(x, y)). \tag{3.189}$$

Then the transverse ("t") derivatives (x and y) are given by

$$\partial_x \{E_m\} = \{\partial_x E_m\} + h_x \{\partial_z E_m\}, \tag{3.190}$$

and

$$\partial_y \{E_m\} = \{\partial_y E_m\} + h_y \{\partial_z E_m\}. \tag{3.191}$$

Using this notation and the continuity of the electric surface current $\vec{K}^e = -\vec{n} \times \vec{H}$, where $\vec{H}$ is the magnetic field, in index form

$$K^e_{2i}(\vec{x}_h) = K^e_{1i}(\vec{x}_h), \tag{3.192}$$

we can write the continuity condition for the normal derivative as[35]

$$\{N_{2i}\} = \mu\{N_{1i}\} + (\varepsilon^{-1} - 1)V_i(\vec{x}_h), \tag{3.193}$$

where V_i can be written in terms of transverse partial derivatives involving the normal components of the electric field as

$$V_i(\vec{x}_h) = n_m \partial_{it}\{\hat{n}_m \hat{n}_j E_{1j}\} - n_i \partial_{qt}\{\hat{n}_{qt}\hat{n}_j E_{1j}\}. \tag{3.194}$$

This V_i term looks awkward, but it can be integrated by parts. First, choose the boundary unknowns as

$$E_{1i}(\vec{x}_h) = \{E_{1i}\} \doteq \{E_i\}, \tag{3.195}$$

and

$$N_{1i}(\vec{x}_h) = \{N_{1i}\} \doteq \{N_i\}. \tag{3.196}$$

Then we can write the equation for the upper region (3.182) as

$$\frac{1}{2}\{E'_i\} + \iint_D [G_1^{(3p)}(\vec{x}'_h, \vec{x}_h)\{N_i\} - N_1^{(3p)}(\vec{x}'_h, \vec{x}_h)\{E_i\}] d\vec{x}_\perp = E_i^{in}(\vec{x}'_h). \tag{3.197}$$

Here $\{E'_i\}$ means the exterior primed variable placed on the surface, i.e. $\vec{x}'_h$. The equation (3.197) is diagonal in the index. The coupling is from the lower equation (3.184) written using (3.193) through (3.196) as

$$\frac{1}{2}C_{ij}(\vec{x}'_h)\{E'_j\} = \iint_D [G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)(\mu\{N_i\} + (\varepsilon^{-1} - 1)V_i(\vec{x}_h)) - N_2^{(3p)}(\vec{x}'_h, \vec{x}_h)C_{ij}(\vec{x}_h)\{E_j\}] d\vec{x}_\perp. \tag{3.198}$$

The V_i term can be integrated by parts to yield

$$\iint_D G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)V_i(\vec{x}_h) d\vec{x}_\perp = \iint_D V_{ij}(\vec{x}'_h, \vec{x}_h)\{E_i\} d\vec{x}_\perp, \tag{3.199}$$

where

$$V_{ij}(\vec{x}'_h, \vec{x}_h) = \partial_{qt}\{n_i G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\}\hat{n}_{qt}\hat{n}_j - \partial_{it}\{n_m G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\}\hat{n}_m\hat{n}_j. \tag{3.200}$$

We can simplify (3.200) to yield

$$V_{ij}(\vec{x}'_h, \vec{x}_h) = N_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\hat{n}_i\hat{n}_j - \{\partial_i G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\}n_j, \tag{3.201}$$

where now the derivative of the Green's function is taken first, and then the result set on the surface. Combining these results we can rewrite (3.198) as

$$\frac{1}{2}C_{ij}(\vec{x}'_h)\{E'_j\} = \iint_D [G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\mu\{N_i\} - W_{ij}(\vec{x}'_h, \vec{x}_h)\{E_j\}] d\vec{x}_\perp, \tag{3.202}$$

where

$$W_{ij}(\vec{x}'_h, \vec{x}_h) = N_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\delta_{ij} + (\varepsilon^{-1} - 1)\{\partial_i G_2^{(3p)}(\vec{x}'_h, \vec{x}_h)\}n_j. \tag{3.203}$$

Note that (3.197) and (3.202), if put in matrix form, are diagonal in three of the four matrix blocks multiplying the six-dimensional vector of boundary unknowns $[\{E_i\},\{N_i\}]^T$ where T is transpose. The only coupling occurs in the single block of the electric fields from (3.202). We note this in contrast to pre-conditioning methods used to sparsify matrix inversion problems. Here the results are exact and highly sparse as formulated. They have been used computationally to treat the scattering from a body of revolution[41].

We can use these representations to write plane wave representations for the scattered and transmitted fields, above and below the largest surface excursions. From (3.177) we can write the scattered field above the highest surface excursion ($z' > max(h)$) as

$$E_i^{sc}(\vec{x}') = -\iint_D [G_1^{(3p)}(\vec{x}',\vec{x}_h)\{N_i\} - N_1^{(3p)}(\vec{x}',\vec{x}_h)\{E_i\}]d\vec{x}_\perp, \tag{3.204}$$

where now the Green's function is, following (3.59), with the field point above the surface

$$G_1^{(3p)}(\vec{x}',\vec{x}_h) = \frac{i}{2k_1L_1L_2}\sum_{j=-\infty}^{\infty}\sum_{j'=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\beta_{j'}(y'-y)+\gamma_{1jj'}(z'-h(\vec{x}_\perp))]}{\gamma_{1jj'}}, \tag{3.205}$$

and

$$N_1^{(3p)}(\vec{x}',\vec{x}_h) = n_q\partial_q G_1^{(3p)}(\vec{x}',\vec{x}_h). \tag{3.206}$$

Combining these results we can write the scattered field exactly above the highest surface excursion as a plane wave expansion in terms of purely up-going waves as

$$E_i^{sc}(\vec{x}') = \sum_{j=-\infty}^{\infty}\sum_{j'=-\infty}^{\infty}A_{ijj'}\exp[ik_1(\alpha_j x'+\beta_{j'}y'+\gamma_{1jj'}z')], \tag{3.207}$$

where

$$A_{ijj'} = \frac{1}{L_1L_2}\iint_D A_{ijj'}(\vec{x}_\perp)\exp[-ik_1(\alpha_j x+\beta_{j'}y+\gamma_{1jj'}h(\vec{x}_\perp))]d\vec{x}_\perp, \tag{3.208}$$

and

$$A_{ijj'}(\vec{x}_\perp) = \frac{-i}{8\pi^2k_1\gamma_{1jj'}}[\{N_i\}+ik_1(\gamma_{1jj'}-\alpha_j h_x-\beta_{j'}h_y)\{E_i\}], \tag{3.209}$$

in terms of the boundary unknowns.

Similarly, from (3.183), we have the transmitted field below the lowest surface excursion ($z' < min(h)$)

$$E_{2i}(\vec{x}') = \iint_D [G_2^{(3p)}(\vec{x}',\vec{x}_h)N_{2i}(\vec{x}_h) - N_2^{(3p)}(\vec{x}',\vec{x}_h)E_{2i}(\vec{x}_h)]d\vec{x}_\perp, \tag{3.210}$$

where now

$$G_2^{(3p)}(\vec{x}',\vec{x}_h) = \frac{i}{8\pi^2k_1L_1L_2}\sum_{j=-\infty}^{\infty}\sum_{j'=-\infty}^{\infty}\frac{\exp[ik_1(\alpha_j(x'-x)+\beta_{j'}(y'-y)-\gamma_{2jj'}(z'-h(\vec{x}_\perp))]}{\gamma_{2jj'}}, \tag{3.211}$$

and

$$N_2^{(3p)}(\vec{x}',\vec{x}_h) = n_q\partial_q G_2^{(3p)}(\vec{x}',\vec{x}_h). \tag{3.212}$$

The result is the plane wave spectral representation for the transmitted field below the lowest surface excursion in terms of purely down-going waves as

$$E_{2i}(\vec{x}') = \sum_{j=-\infty}^{\infty} \sum_{j'=-\infty}^{\infty} B_{ijj'} \exp[ik_1(\alpha_j x' + \beta_{j'} y' - \gamma_{2jj'} z')], \tag{3.213}$$

where

$$B_{ijj'} = \frac{1}{L_1 L_2} \iint_D B_{ijj'}(\vec{x}_\perp) \exp[-ik_1(\alpha_j x + \beta_{j'} y - \gamma_{2jj'} h(\vec{x}_\perp))] d\vec{x}_\perp, \tag{3.214}$$

and

$$B_{ijj'}(\vec{x}_\perp) = \frac{i}{8\pi^2 k_1 \gamma_{2jj'}} [N_{2i}(\vec{x}_h) - ik_1(\gamma_{2jj'} + \alpha_j h_x + \beta_{j'} h_y) E_{2i}(\vec{x}_h)], \tag{3.215}$$

written in terms of the boundary values from the lower region. Using the boundary conditions (3.187) and (3.193) and integration by parts we can rewrite (3.215) in terms of the boundary unknowns as

$$B_{ijj'}(\vec{x}_\perp) = \frac{i}{8\pi^2 k_1 \gamma_{2jj'}} [\mu \{N_i\} - W_{ijj'l}(\vec{x}_\perp)\{E_l\}], \tag{3.216}$$

where

$$W_{ijj'l}(\vec{x}_\perp) = ik_1[\alpha_j h_x + \beta_{j'} h_y + \gamma_{2jj'})\delta_{il} - (\varepsilon^{-1} - 1)(\delta_{i1}\alpha_j + \delta_{i2}\beta_{j'} - \delta_{i3}\gamma_{2jj'})n_l]. \tag{3.217}$$

Equations (3.207) and (3.213) are the exact plane wave representations in the appropriate regions. In the next section we write general partial spectral representations of the fields, and show the relations between them and the plane wave spectral representations here which are valid in limited domains.

3.10 Partial Spectral Methods for Electromagnetic Problems

We develop this section in analogy with the scalar results in Sec.5. This is the electromagnetic version of the Spectral-Coordinate approach. We define the three-dimensional plane wave states in the upper region 1 for up$^{(+)}$- and down$^{(-)}$-going waves as

$$\phi^{\pm}_{1jj'}(\vec{x}) = \exp[ik_1(-\alpha_j x - \beta_{j'} y \pm \gamma_{1jj'} z), \tag{3.218}$$

where $\gamma_{1jj'}$ is defined following (3.59). The function satisfies the three-dimensional Helmholtz equation

$$(\nabla_3^2 + k_1^2)\phi^{\pm}_{1jj'}(\vec{x}) = 0. \tag{3.219}$$

The incident electric field can be written as a general plane wave expansion of down-going waves

$$E_i^{in}(\vec{x}) = \sum_{j=-\infty}^{\infty} \sum_{j'=-\infty}^{\infty} I_{ijj'} \exp[ik_1(\alpha_j x + \beta_{j'} y - \gamma_{1jj'} z)]. \tag{3.220}$$

Apply Green's theorem in the domain defined by Θ_1 in (3.176) to $\phi^{\pm}_{1jj'}$ and E_{1i}, use the two-dimensional Floquet conditions to cancel the side integrals as in Sec.9 and the result is

$$\frac{1}{L_1 L_2} \iint_D \phi^{\pm}_{1jj'}(\vec{x}_h) U^{\pm}_{ijj'}(\vec{x}_h) d\vec{x}_\perp = \gamma_{1jj'} \left\{ \begin{matrix} -I_{ijj'} \\ A_{ijj'} \end{matrix} \right\}, \tag{3.221}$$

where U is defined as

$$U^{\pm}_{ijj'}(\vec{x}_h) = \frac{1}{2ik_1}[\{N_i\} - ik_1(\pm\gamma_{1jj'} + \alpha_j h_x + \beta_{j'} h_y)\{E_i\}]. \tag{3.222}$$

The $A_{ijj'}$ are the spectral coefficients of the scattered field from (3.207). Recall that the scattered field is evaluated on a flat surface ($z = H_1$) above the highest surface excursion, so the representation (3.207) is rigorously valid and not a Rayleigh approximation. We also used the boundary values (3.195) and (3.196). In (3.221), the up-going plane wave states ϕ^+ project out the down-going incident field spectral components $I_{ijj'}$, and the down-going plane wave states ϕ^- project out the up-going scattered field spectral components $A_{ijj'}$. Equations (3.221) and (3.222) are the vector generalizations of (3.105) and (3.106).

For the lower region 2, the three-dimensional up- and down-going plane wave states are defined as

$$\phi^{\pm}_{2jj'}(\vec{x}) = \exp[ik_1(-\alpha_j x - \beta_{j'} y \pm \gamma_{2jj'} z)], \tag{3.223}$$

where $\gamma_{2jj'}$ is defined following (3.62). The functions satisfy the three-dimensional Helmholtz equation

$$(\nabla_3^2 + k_2^2)\phi^{\pm}_{2jj'}(\vec{x}) = 0. \tag{3.224}$$

Green's theorem on $\phi^{\pm}_{2jj'}$ and the total transmitted field E_{2i} in the domain defined by $\Theta_2(\vec{x}) = 1 - \Theta_1(\vec{x})$ yields the relations

$$\frac{1}{L_1 L_2}\iint_D \phi^{\pm}_{2jj'}(\vec{x}_h) L^{\pm}_{ijj'}(\vec{x}_h) d\vec{x}_{\perp} = -\gamma_{2jj'}\left\{{}^{B_{ijj'}}_{0}\right\}, \tag{3.225}$$

where

$$L^{\pm}_{ijj'}(\vec{x}_h) = \frac{1}{2ik_1}[N_{2i}(\vec{x}_h) - ik_1(\pm\gamma_{2jj'} + \alpha_j h_x + \beta_{j'} h_y) E_{2i}(\vec{x}_h)], \tag{3.226}$$

in terms of the boundary values from the lower region. Using the boundary values (3.195) and (3.196) and integration by parts, (3.226) can be rewritten as

$$L^{\pm}_{ijj'}(\vec{x}_h) = \frac{1}{2ik_1}[\mu\{N_i\} - W^{\pm}_{ijj'l}(\vec{x}_h)\{E_l\}], \tag{3.227}$$

with a sum over $l = (1,2,3)$ and where

$$W^{\pm}_{ijj'l}(\vec{x}_h) = ik_1[(\alpha_j h_x + \beta_{j'} h_y \pm \gamma_{2jj'})\delta_{il} - (\varepsilon^{-1} - 1)(\alpha_j \delta_{i1} + \beta_{j'}\delta_{i2} \mp \gamma_{2jj'}\delta_{i3})n_l]. \tag{3.228}$$

Note that $W^{+}_{ijj'l}$ is just $W_{ijj'l}$ from (3.217). Equations (3.225) and (3.227) are the vector generalizations of (3.111) and (3.112). The procedure is to solve the upper equation (3.221) and the lower equation (3.225) for the boundary unknowns $\{Ni\}$ and $\{E_i\}$ and evaluate the remaining equations for the scattered and transmitted amplitudes.

3.11 Full Spectral Methods for Electromagnetic Problems

In this section we develop the full spectral methods using the conjugate Rayleigh basis in analogy with Sec.8 for the scalar case. In (3.221) and (3.225) we use the following expansions in the conjugate Rayleigh basis,

$$\{E_i\} = \sum_{l=-\infty}^{\infty}\sum_{l'=-\infty}^{\infty} E_{ill'}\bar{\phi}^{+}_{1ll'}(\vec{x}_h), \tag{3.229}$$

and

$$\{N_i\} = ik_1 \sum_{l=-\infty}^{\infty} \sum_{l'=-\infty}^{\infty} N_{ill'} \bar{\phi}^+_{1ll'}(\vec{x}_h), \tag{3.230}$$

with the normal derivative on the boundary scaled by ik_1 and where $\phi^+_{1ll'}$ is from (3.218). The overbar is complex conjugation. Integrate the slope terms by parts as for example

$$< \phi^{\pm}_{1jj'}, h_x \phi^+_{1ll'} >= \frac{\alpha_j - \alpha_l}{\pm\gamma_{1jj'} - \bar{\gamma}_{1ll'}} < \phi^{\pm}_{1jj'}, \phi^+_{1ll'} >, \tag{3.231}$$

and the equations for the upper region can be written using (3.221) as

$$\sum_{l=-\infty}^{\infty} \sum_{l'=-\infty}^{\infty} < \phi^{\pm}_{1jj'}, \phi^+_{1ll'} > [N_{ill'} - U^{\pm}(jj',ll')E_{ill'}] = 2\gamma_{1jj'} \left\{ \begin{matrix} -I_{ijj'} \\ A_{ijj'} \end{matrix} \right\}, \tag{3.232}$$

where

$$U^{\pm}(jj',ll') = \frac{1 - \alpha_j\alpha_l - \beta_{j'}\beta_{l'} \mp \gamma_{1jj'}\bar{\gamma}_{1ll'}}{\pm\gamma_{1jj'} - \bar{\gamma}_{1ll'}}. \tag{3.233}$$

We have written the double spectral values jj' and ll' as arguments of U in illustration of the fact that they are each replacing coordinate sampling/integration along two-dimensional surfaces denoted by $\vec{x}'_h$ and $\vec{x}_h$ respectively, as well as to indicate that the equations (3.232) are diagonal in the vector index $''i''$. That is, the i^{th} component of A is related to the i^{th} components of N and E. There is no coupling in this index for the equations from region 1. It can be shown that the matrix $< \phi^+_{1jj'}, \phi^+_{1ll'} >$ is self-adjoint, positive definite and hence invertible.

For the lower region 2 these same expansions and integration of the slope terms yields from (3.225)

$$\sum_{l=-\infty}^{\infty} \sum_{l'=-\infty}^{\infty} < \phi^{\pm}_{2jj'}, \phi^+_{1ll'} > [\mu N_{ill'} - L^{\pm}_{ip}(jj',ll')E_{pll'}] = -2\gamma_{2jj'} \left\{ \begin{matrix} B_{ijj'} \\ 0 \end{matrix} \right\}, \tag{3.234}$$

where there is an implicit sum over the repeated subscript $p = (1,2,3)$. The full coupling of these equations resides in this summation. Here the L term can be written as

$$L^{\pm}_{ip}(jj',ll') = \frac{M^{\pm}_{ip}(jj',ll')}{\pm\gamma_{2jj'} - \bar{\gamma}_{1ll'}}, \tag{3.235}$$

where M can be written as a diagonal (D) part and a full (F) part, the latter of which contains the coupling,

$$M^{\pm}_{ip}(jj',ll') = D^{\pm}(jj',ll')\delta_{ip} + F^{\pm}_{ip}(jj',ll'), \tag{3.236}$$

where

$$D^{\pm}(jj',ll') = K^2 - \alpha_j\alpha_l - \beta_{j'}\beta_{l'} \mp \gamma_{2jj'}\bar{\gamma}_{1ll'}, \tag{3.237}$$

and

$$F^{\pm}_{ip}(jj',ll') = (\varepsilon^{-1} - 1)(\alpha_j\delta_{i1} + \beta_{j'}\delta_{i2} \mp \gamma_{2jj'}\delta_{i3})[(\alpha_j - \alpha_l)\delta_{p1} + (\beta_{j'} - \beta_{l'})\delta_{p2} - (\pm\gamma_{2jj'} - \bar{\gamma}_{1ll'})\delta_{p3}]. \tag{3.238}$$

The procedure is to solve the upper equation (3.232) and the lower equation (3.234), which is a spectral extinction equation, for the unknown expansion coefficients $N_{ill'}$ and $E_{ill'}$, and evaluate the remaining equations for the scattered $A_{ijj'}$ and transmitted $B_{ijj'}$ spectral coefficients.

3.12 Summary

We have derived exact formal sets of equations, in both coordinate and various spectral domains, to describe the scattering from deterministic gratings. Both acoustic scalar one-dimensional problems and full electromagnetic two-dimensional problems were considered. Both involved a grating surface separating two homogeneous regions of space. Both involved coordinate-space representations from which proceeded rigorous plane wave spectral representations valid for the scattered field above the highest surface excursion and for the transmitted field below its lowest excursion. The electromagnetic development was treated in analogy with the scalar problem, with boundary conditions derived for the electric field and its normal derivative from the standard boundary conditions on currents and the normal components of the displacement vector.

From these coordinate representations we proceeded first to partial spectral representations where the word "partial" refers to the field variables. These could be derived in a straightforward way just using plane waves and Green's theorem, and without involving the Green's function explicitly. We stress again that the equations are exact. In addition, these led to surface inversion examples for the scalar case using perturbation theory (where the boundary values could be factored out), and the Kirchhoff approximation (where the boundary values were approximated).

The full spectral equations involved expanding the boundary unknowns in some set of functions, and it is here where the Rayleigh and Waterman assumptions come into play. For the scalar case we presented three expansions. The first was a physical optics modified Fourier expansion with a single plane wave modulating the surface fields. The second was what we referred to as a Floquet-Fourier basis which modulated the Fourier expansion by still preserving the Floquet-periodicity of the surface fields but without the full plane waves, and the third was an expansion in the conjugate Rayleigh basis where each term in the expansion could be thought of as modulated by a plane wave. For the electromagnetic case only the expansion in the conjugate Rayleigh basis was considered. Since we used a scalar analogy for the electromagnetic problem the resulting equations were formally analogous to the scalar equations with the additional complication being first a vector problem, and second the Bragg modal sampling in two two-dimensional spaces, the spaces of boundary and field points.

We pointed out in the paper where any of these equations have been solved, but we repeat that the full computational results and the comparisons of different computational results for this problem require at least a separate paper if not a separate book.

Appendix 3.A. A Note on Matrix Elements

For the fully spectral methods in Secs.7 and 8 for the acoustic case and Sec.11 for the electromagnetic case, the matrix elements have a characteristic form. In the one-dimensional case, after projection on the various basis sets considered in this paper, (see (3.158), (3.165), (3.170), and (3.174)), they have the general form

$$M(a,b) = \frac{1}{L}\int_{-L/2}^{L/2} \exp[iax + ik_1 bh(x)]dx. \tag{3.239}$$

For all the cases in question, $a = 2\pi(j'-j)/L$ which is the Fourier part common to all, and $b = \pm\gamma_{pj} - \gamma_{10}$ for the physical optics case with one overall plane wave, $b = \pm\gamma_{pj}$ for the Floquet-Fourier case with no plane waves modulating the field expansion, and $b = \pm\gamma_{pj} - \bar{\gamma}_{1j'}$ for the conjugate Rayleigh case with plane waves related to each Bragg mode in the sum. They have a general validity in surface scattering problems due to the presence of Green's functions or plane wave type expansions. For example, for a random surface, these functions M were referred to as interaction functions in a Feynman diagram expansion[114,29,30], essentially a perturbation expansion in the functions.

In addition, for many surfaces, not necessarily analytic ones, the integral can be expressed in closed form in terms of special functions. For example, a cosine surface yields Bessel functions for M, a symmetric sawtooth function yields simple exponentials, a quadratic surface yields Fresnel integrals, a vortex-like surface involving a logarithm yields cosine integrals which can be evaluated in closed form (or, in a different form, confluent hypergeometric functions), a cycloid can be evaluated in terms of Bessel functions, a full-wave rectified surface in terms of a Bessel series, and a periodic array of semicircular cylinders (bosses)[98] in terms of a Bessel series. These closed form solutions can be useful in computations or for approximations. The details can be found in[32]. Two-dimensional integrals occurring in the electromagnetic problem can be developed in a similar way for egg-crate surfaces of the form $h(x,y) = h_1(x) + h_2(y)$.

References:

[1] T.Abboud, and H.Ammari, "Diffraction at a Curved Grating: Approximation by an Infinite Plane Grating", J.Math.Anal.Appl. **202**, 1076-1100(1996).

[2] H.D.Alber, "A Quasi-periodic Boundary Value Problem for the Laplacian and the Continuation of its Resolvent", Proc.Roy.Soc.Edinburgh **82A**, 251-272(1979).

[3] T.Arens, S.N.Chandler-Wilde, and J.A.DeSanto, "On Integral Equation and Least Squares Methods for Scattering by Diffraction Gratings", Comm.Comp.Phys. **1**, 1010-1042(2006).

[4] M.Bagieu, and D.Maystre, "Waterman and Rayleigh Methods for Diffraction Grating Problems: Extension of the Convergence Domain", J.Opt.Soc.Am. A **15**, 1566-1576(1998).

[5] M.Bagieu, and D.Maystre, "Regularized Waterman and Rayleigh Methods: Extension to Two-dimensional Gratings", J.Opt.Soc.Am. A **16**, 284-292(1999).

[6] G.Bao, D.C.Dobson, and J.A.Cox, "Mathematical Studies in Rigorous Grating Theory", J.Opt.Soc.Am. A **12**, 1029-1042(1995).

[7] G.Bao, "Diffraction by a Periodic Surface", In: *Mathematical and Numerical Aspects of Wave Propagation*, Ed. J.A.DeSanto (SIAM Publications, Philadelphia, 1998) pp. 476-478.

[8] R.G.Barantsev, "Plane Wave Scattering by a Double Periodic Surface of Arbitrary Shape", Soviet Physics-Acoustics **7**, 123-126(1961).

[9] G.R.Barnard, C.W.Horton, M.K.Miller, and F.R.Spitznogle, "Underwater-Sound Reflection from a Pressure-Release Sinusoidal Surface", J.Acoust.Soc.Am. **39**, 1162-1169(1966).

[10] R.H.T.Bates, "Analytic Constraints on Electromagnetic Field Computations", IEEE Trans. **MTT-23**, 605-623(1975).

[11] P.M.van den Berg, and J.T.Fokkema, "The Rayleigh Hypothesis in the Theory of Reflection by a Grating", J.Opt.Soc.Am. **69**, 27-31(1979).

[12] P.M.van den Berg, "Reflection by a Grating: Rayleigh Methods", J.Opt.Soc.Am. **71**, 1224-1229(1981).

[13] P.M.van den Berg, "Smith-Purcell Radiation from a Line Charge Moving Parallel to a Reflection Grating", J.Opt.Soc.Am **63**, 689-698(1973).

[14] D.H.Berman, and J.S.Perkins, "Rayleigh Method for Scattering from Random and Deterministic Interfaces", J.Acoust.Soc.Am. **88**, 1032-1044(1990).

[15] G.C.Bishop, and J.Smith, "A Scattering Model for Nondifferentiable Periodic Surface Roughness", J.Acoust.Soc.Am. **91**, 744-770(1991).

[16] L.C.Botten, "A New Formalism for Transmission Gratings", Opt.Acta. **25**, 481-499(1978).

[17] G.Bruckner, and J.Elschner, "A Two-Step Algorithm for the Reconstruction of Perfectly Reflecting Periodic Profiles", Inv.Pbs. **19**, 315-329(2003).

[18] O.P.Bruno, and F.Reitich, "Numerical Solution of Diffraction Problems:A Method of Variation of Boundaries", J.Opt.Soc.Am.A **10**, 1168-1175(1993).

[19] O.P.Bruno, and F.Reitich, "Numerical Solution of Diffraction Problems: A Method of Variation of Boundaries. II. Finitely Conducting Gratings, Pad*é* Approximants, and Singularities", J.Opt.Soc.Am.A **10**, 2307-2316(1993).

[20] O.P.Bruno, and F.Reitich, "Numerical Solution of Diffraction Problems: A Method of Variation of Boundaries. III. Doubly Periodic Gratings", J.Opt.Soc.Am. A **10**, 2551-2562(1993).

[21] O.P.Bruno, and M.C.Haslam, "Efficient High-Order Evaluation of Scattering by Periodic Surfaces: Deep Gratings, High Frequencies, and Glancing Incidences", J.Opt.Soc.Am. A **26**, 658-668(2009).

[22] O.P.Bruno, and M.C.Haslam, "Efficient High-order Evaluation of Scattering by Periodic Surfaces: Vector-parametric Gratings and Geometric Singularities", Waves Random Complex Media **20**, 530-550(2010).

[23] J-M.Chesneaux, and A.Wirgin, "Reflection from a Corrugated Surface Revisited", J.Acoust.Soc.Am. **96**, 1116-1129(1994).

[24] J-M.Chesneaux, and A.Wirgin, "Response to 'Comments on 'Reflection from a Corrugated Surface Revisited' ", J.Acoust.Soc.Am. **98**, 1815-1816(1995).

[25] S.Christiansen, and R.E.Kleinman, "On a Misconception Involving Point Collocation and the Rayleigh Hypothesis", IEEE Trans. **AP-44**, 1309-1316(1996).

[26] S-L.Chuang, and J.A.Kong, "Scattering of Waves from Periodic Surfaces", Proc.IEEE **69**, 1132-1144(1981).

[27] S-L.Chuang, and J.A.Kong, "Wave Scattering from a Periodic Dielectric Surface for a General Angle of Incidence", Rad.Sci. **17**, 545-557(1982).

[28] P.C.Clemmow, *The Plane Wave Spectrum Representation of Electromagnetic Fields*, International Series of Monographs in Electromagnetic Waves, **12**, (Pergamon, Oxford, 1966).

[29] J.A.DeSanto, "Scattering from a Random Rough Surface: Diagram Methods for Elastic Media", J.Math.Phys. **14**, 1566-1573(1973).

[30] J.A.DeSanto, "Green's Functions for Electromagnetic Scattering from a Random Rough Surface", J.Math.Phys. **15**, 283-288(1974).

[31] J.A.DeSanto, "Scattering from a Sinusoid: Derivation of Linear Equations for the Field Amplitudes", J.Acoust.Soc.Am. **57**, 1195-1197(1975).

[32] J.A.DeSanto, "Scattering from a Perfectly Reflecting Arbitrary Periodic Surface: An Exact Theory", Radio Sci. **16**, 1315-1326(1981).

[33] J.A.DeSanto, "Exact Spectral Formalism for Rough-Surface Scattering", J.Opt.Soc.Am. A **2**, 2202-2207(1985).

[34] J.A.DeSanto, *Scalar Wave Theory*, (Springer, Berlin, Heidelberg, New York, 1992).

[35] J.A.DeSanto, "A new Formulation of Electromagnetic Scattering from Rough Dielectric Interfaces", J.Elect.Waves Appl. **7**, 1793-1806(1993).

[36] J.A.DeSanto, G.Erdmann, W.Hereman, and M.Misra, "Theoretical and Computational Aspects of Scattering from Rough Surfaces: One-dimensional Perfectly Reflecting Surfaces", Waves Random Media **8**, 385-414(1998).

[37] J.A.DeSanto, G.Erdmann, W.Hereman, and M.Misra, "Theoretical and Computational Aspects of Scattering from Periodic Surfaces: One-dimensional Transmission Interface", Waves Random Media **11**, 425-453(2001).

[38] J.A.DeSanto, G.Erdmann, W.Hereman, B.Krause, M.Misra, and E.Swim, "Theoretical and Computational Aspects of Scattering from Periodic Surfaces: Two-dimensional Perfectly Reflecting Surfaces Using the Spectral-Coordinate Method", Waves Random Media **11**, 455-487(2001).

[39] J.A.DeSanto, G.Erdmann, W.Hereman, B.Krause, M.Misra, and E.Swim, "Theoretical and Computational Aspects of Scattering from Periodic Surfaces: Two-dimensional Transmission Surfaces Using the Spectral-Coordinate Method", Waves Random Media **11**, 489-526(2001).

[40] J.A.DeSanto, "Scattering by Rough Surfaces", In: *Scattering*, Eds. R.Pike, P.Sabatier (Academic, New York, 2002) pp. 15-36.

[41] J.A.DeSanto, and A.Yuffa, "A New Integral Equation Method for Direct Electromagnetic Scattering in Homogeneous Media and Its Numerical Confirmation", Waves Random Complex Media **16**, 397-408(2006).

[42] J.Elschner, and G.Schmidt, "Diffraction in Periodic Structures and Optimal Design of Binary Gratings I. Direct Problems and Gradient Formulas", Math.Meth.Appl.Sci. **21**, 1297-1342(1998).

[43] B.Gallinet, A.M.Kern, and O.J.F.Martin, "Accurate and Versatile Modeling of Electromagnetic Scattering on Periodic Nanostructures with a Surface Integral Approach", J.Opt.Soc.Am. A **27**, 2261-2271(2010).

[44] N.Garcia, and N.Cabrera, "New Method for Solving the Scattering of Waves from a Periodic Hard Surface: Solutions and Numerical Comparisons with the Various Formalisms", Phys.Rev. B **18**, 576-589(1978).

[45] N.Garcia, V.Celli, N.R.Hill, and N.Cabrera, "Ill-conditioned Matrices in the Scattering of Waves from Hard Corrugated Surfaces", Phys.Rev. B **18**, 5184-5189(1978).

[46] N.E.Glass, and A.A.Maradudin, "Surface Plasmons on a Large-Amplitude Grating", Phys.Rev. B **24**, 595-602(1981).

[47] N.E.Glass, A.A.Maradudin, and V.Celli, "Surface Plasmons on a Large-amplitude Doubly Periodically Corrugated Surface", Phys.Rev. B **26**, 5357-5365(1982).

[48] N.R.Hill, and V.Celli, "Limits of Convergence of the Rayleigh Method for Surface Scattering", Phys.Rev. B **17**, 2478-2481(1978).

[49] R.L.Holford, "Scattering of Sound Waves at a Periodic, Pressure-release Surface: An Exact Solution", J.Acoust.Soc.Am. **70**, 1116-1128(1981).

[50] J.P.Hugonin, R.Petit, and M.Cadilhac, "Plane-wave Expansions Used to Describe the Field Diffracted by a Grating", J.Opt.Soc.Am. A **71**, 593-598(1981).

[51] M.C.Hutley, and M.V.Bird, "A Detailed Experimental Study of the Anomalies of a Sinusoidal Diffraction Grating", Optica Acta **20**, 771-782(1973).

[52] M.C.Hutley, *Diffraction Gratings*, Techniques of Physics **6**, (Academic, London, 1982).

[53] H.Ikuno, and K.Yasuura, "Improved Point-Matching Method with Application to Scattering from a Periodic Surface", IEEE Trans. **AP-21**, 657-662(1973).

[54] V.Jamnejad-Dailami, R.Mittra, and T.Itoh, "A Comparative Study of the Rayleigh Hypothesis and Analytic Continuation Methods as Applied to Sinusoidal Gratings", IEEE Trans. **AP-20**, 392-394(1972).

[55] A.K.Jordan, and R.H.Lang, "Electromagnetic Scattering Patterns from Sinusoidal Surfaces", Rad.Sci. **14**, 1077-1088(1979).

[56] L.Kazandjian, "Comparison of the Rayleigh-Fourier and Extinction Theorem Methods Applied to Acoustic Scattering in a Waveguide", J.Acoust.Soc.Am. **90**, 2623-2627(1991).

[57] J.B.Keller, "Singularities and Rayleigh's Hypothesis for Diffraction Gratings", J.Opt.Soc.Am. A **17**, 456-457(2000).

[58] M.-J.Kim, H.M.Berenyi, and R.E.Burge, "Scattering of Scalar Waves by Two-dimensional Gratings of Arbitrary Shape: Application to Rough Surfaces at Near-grazing Incidence", Proc.R.Soc.Lond. A **446**, 298-308(1994).

[59] A.Kirsch, "Uniqueness Theorems in Inverse Scattering Theory for Periodic Structures", Inv.Pbs. **10**. 145-152(1994).

[60] A.I.Kleev, and A.B.Manenkov, "The Convergence of Point-Matching Techniques", IEEE Trans. **AP-37**, 50-54(1989).

[61] J.A.Kong, *Electromagnetic Wave Theory*, (Wiley, New York, 1986).

[62] A.Lakhtakia, V.K.Varadan, and V.V.Varadan, "On the Acoustic Response of a Deeply Corrugated Periodic Surface-A Hybrid T-matrix Approach", J.Acoust.Soc.Am. **78**, 2100-2104(1985).

[63] B.Laks, D.L.Mills, and A.A.Maradudin, "Surface Polaritons on Large-amplitude Gratings", Phys.Rev. B **23**, 4965-4976(1981).

[64] B.A.Lippmann, "Note on the Theory of Gratings", J.Opt.Soc.Am. **43**, 408(1953).

[65] E.G.Loewen, M.Nevière, and D.Maystre, "On an Asymptotic Theory of Diffraction Gratings Used in the Scalar Domain", J.Opt.Soc.Am. **68**, 496-502(1978).

[66] H.W.Marsh, "In Defense of Rayleigh's Scattering from Corrugated Surfaces", J.Acoust.Soc.Am. **35**, 1835-1836(1963).

[67] R.I.Masel, R.P.Merrill, and W.H.Miller, "Quantum Scattering from a Sinusoidal Hard Wall: Atomic Diffraction from Solid Surfaces", Phys.Rev. B **12**, 5545-5551(1975).

[68] D.Maystre, "A New General Theory for Dielectric Coated Gratings", J.Opt.Soc.Am. **68**, 490-495(1978).

[69] D.Maystre, "Rigorous Vector Theories of Diffraction Gratings", In: Progress In Optics XXI, Ed. E.Wolf, (Elsevier, New York, 1984) pp.1-67.

[70] D.Maystre, and M.Cadilhac, "Singularities of the Continuation of Fields and Validity of Rayleigh's Hypothesis", J.Math.Phys. **26**, 2201-2204(1985).

[71] D.F.McCammon, and S.T.McDaniel, "Surface Velocity, Shadowing, Multiple Scattering, and Curvature on a Sinusoid", J.Acoust.Soc.Am. **79**, 1778-1785(1986).

[72] R.C.McNamara, and J.A.DeSanto, "Numerical Determination of Scattered Field Amplitudes for Rough Surfaces", J.Acoust.Soc.Am. **100**, 3519-3526(1996).

[73] P.E.McSharry, D.T.Moroney, and P.J.Cullen, "Wave Scattering by a Two-dimensional Pressure Release Surface Based on a Perturbation of the Green's Function", J.Acoust.Soc.Am. **98**, 1699-1716(1995).

[74] W.C.Meecham, "Variational Method for the Calculation of the Distribution of Energy Reflected from a Periodic Surface. I.", J.Appl.Phys. **27**, 361-367(1956).

[75] W.C.Meecham, "Point Source Transmission Through a Sinusoidal Ocean Surface", J.Acoust.Soc.Am. **64**, 1478-1481(1978).

[76] R.F.Millar, "On the Rayleigh Assumption in Scattering by a Periodic Surface", Proc.Camb.Phil.Soc. **65**, 773-791(1969).

[77] R.F.Millar, "The Location of Singularities of Two-dimensional Harmonic Functions. I. Theory", SIAM J.Math.Anal. **1**, 333-344(1970).

[78] R.F.Millar, "The Location of Singularities of Two-dimensional Harmonic Functions. II. Applications", SIAM J.Math.Anal. **1**, 345-353(1970).

[79] R.F.Millar, "Singularities of Two-dimensional Exterior Solutions of the Helmholtz Equation", Proc.Camb.Phil.Soc. **69**, 175-188(1971).

[80] R.F.Millar, "On the Rayleigh Assumption in Scattering by a Periodic Surface. II", Proc.Camb.Phil.Soc. **69**, 217-225(1971).

[81] R.F.Millar, "The Rayleigh Hypothesis and a Related Least-squares Solution to Scattering Problems for Periodic Surfaces and Other Scatterers, Rad. Sci. **8**, 785-796(1973).

[82] R.F.Millar, Singularities and the Rayleigh Hypothesis for Solutions to the Helmholtz Equation", IMA J.Appl.Math. **37**, 155-171(1986).

[83] J.Nakayama, L.Gao, and Y.Tamura, "Scattering of a Plane Wave from a Periodic Random Surface: A Probabilistic Approach", Waves Random Media **7**, 65-78(1997).

[84] J.Nakayama, and Y.Kitada, "Wave Scattering from a Finite Periodic Surface: Spectral Formalism for TE Wave", ICICE Trans. Electron. **E96-C**, 1098-1105(2003).

[85] M.Neviere, M.Cadilhac, and R.Petit, "Applications of Conformal Mappings to the Diffraction of Electromagnetic Waves by a Grating", IEEE Trans. **AP-21**, 37-46(1973).

[86] D.P.Nichols, and F.Reitich, "Boundary Perturbation Methods for High-Frequency Acoustic Scattering: Shallow Periodic Gratings", J.Acoust.Soc.Am. **123**, 2531-2541(2008).

[87] Y.Okuno, and T.Matsuda, "Mode-Matching Method with a Higher-Order Smoothing Procedure for the Numerical Solution of Diffraction by a Grating", J.Opt.Soc.Am. **73**, 1305-1311(1983).

[88] T.C.Paulick, "Applicability of the Rayleigh Hypothesis to Real Materials", Phys.Rev. B **42**, 2801-2824(1990).

[89] R.Petit, "Electromagnetic Grating Theories: Limitations and Successes", Nouv.Rev.Optique **t.6, no.3**, 129-135(1975).

[90] R.Petit (Ed.) *Electromagnetic Theory of Gratings*, Topics in Current Physics **22**, (Springer, Berlin, 1980).

[91] E.Popov, "Light Diffraction by Relief Gratings: A Macroscopic and Microscopic View", In: Progress In Optics XXXI, Ed. E.Wolf, (Elsevier, New York, 1993) pp.139-187.

[92] E.Popov, and M.Neviere, "Grating Theory: New Equations in Fourier Space Leading to Fast Converging Results for TM Polarization", J.Opt.Soc.Am. A **17**, 1773-1784(2000).

[93] M.Saillard, and J.A.DeSanto, "A Coordinate-Spectral Method for Rough Surface Scattering", Waves Random Media **6**, 135-150(1996).

[94] W.A.Schlup, "On the Convergence of the Rayleigh Ansatz for Hard-Wall Scattering on Arbitrary Periodic Surface Profiles", J.Phys. A: Math.Gen. **17**, 2607-2619(1984).

[95] L.Schwartz, *Mathematics for the Physical Sciences*, (Addison-Wesley, Reading, Massachusetts, 1966).

[96] J.M.Soto-Crespo, and M.Nieto-Vesperinas, "Electromagnetic Scattering from Very Rough Random Surfaces and Deep Reflection Gratings", J.Opt.Soc.Am. A **6**, 367-384(1989).

[97] F.Toigo, A.Marvin, V.Celli, and N.R.Hill, "Optical Properties of Rough Surfaces: General Theory and the Small Roughness Limit", Phys.Rev. B **15**, 5618-5626(1977).

[98] V.Twersky, "On the Scattering of Waves by an Infinite Grating", IRE Trans. **AP-4**, 330-345(1956).

[99] J.L.Uretsky, "Reflection of a Plane Sound Wave from a Sinusoidal Surface", J.Acoust.Soc.Am. **35**, 1293-1294(1963).

[100] J.L.Uretsky, "The Scattering of Plane Waves from Periodic Surfaces", Annals of Physics **33**, 400-427(1965).

[101] G.Valerio, P.Baccarelli, P.Burghignoli, and A.Galli, "Comparative Analysis of Acceleration Techniques for 2-D and 3-D Green's Functions in Periodic Structures Along One and Two Directions", IEEE Trans. **AP-55**,1630-1642(2007).

[102] M.E.Veysoglu, H.T.Yueh, R.T.Shin, and J.A.Kong, "Polarimetric Passive Remote Sensing of Periodic Surfaces", J.Electro.Waves Appl. **5**, 267-280(1991).

[103] A.G.Voronovich, *Wave Scattering From Rough Surfaces*, 2nd edn.(Springer, New York, 1999).

[104] A.G.Voronovich, "Rayleigh Hypothesis", In: *Light Scattering and Nanoscale Surface Roughness*, Ed. A.A.Maradudin, (Springer, New York, 2007) pp. 93-105.

[105] P.C.Waterman, "Scattering by Periodic Surfaces", J.Acoust.Soc.Am. **57**, 791-802(1975).

[106] G.Whitman, and F.Schwering, "Scattering by Periodic Metal Surfaces with Sinusoidal Height Profiles-A Theoretical Approach", IEEE Trans. **AP-25**, 869-876(1997).

[107] G.M.Whitman, D.M.Leskiw, and F.Schwering, "Rigorous Theory of Scattering by Perfectly Conducting Periodic Surfaces with Trapezoidal Height Profile. TE and TM Polarization", J.Opt.Soc.Am. **70**, 1495-1503(1980).

[108] C.H.Wilcox, *Scattering Theory for Diffraction Gratings*, Applied Mathematical Sciences, No.46, (Springer, New York, 1984).

[109] A.Wirgin, "Reflection from a Corrugated Surface", J.Acoust.Soc.Am. **68**, 692-699(1980).

[110] A.Wirgin, "Scattering from Sinusoidal Gratings: An Evaluation of the Kirchhoff Approximation", J.Opt.Soc.Am. **73**, 1028-1041(1983).

[111] R.J.Wombell, and J.A.DeSanto, "The Reconstruction of Shallow Rough-surface Profiles from Scattered Field Data", Inv.Pbs. **7**, L7-L12(1991).

[112] R.J.Wombell, and J.A.DeSanto, "Reconstruction of Rough-surface Profiles with the Kirchhoff Approximation", J.Opt.Soc.Am. A **8**, 1892-1897(1991).

[113] K.A.Zaki, and A.R.Neureuther, "Scattering from a Perfectly Conducting Surface with a Sinusoidal Height Profile: TM Polarization", IEEE Trans. **AP-19**, 747-751(1971).

[114] G.Zipfel, and J.A.DeSanto, "Scattering of a Scalar Wave from a Random Rough Surface: A Diagrammatic Approach", J.Math.Phys. **13**, 1903-1911(1972).

Chapter 4:

Integral Method for Gratings

Daniel Maystre and Evgeny Popov

Table of Contents:

Chapter 4

Integral Method for Gratings

Daniel Maystre and Evgeny Popov

Aix-Marseille Université, CNRS, Centrale Marseille, Institut Fresnel UMR 7249,
Campus de Saint Jerome,13013 Marseille, France
Daniel.maystre@fresnel.fr

4.1. Introduction

Integral methods for scattering problems represent a class of mathematical methods based on integral equations. In this chapter, all the integral equations are deduced from boundary value problems of scattering and are classified as Fredholm integral equations. They can be written in the form:

$$c\, u(\vec{r}) = v(\vec{r}) + \int_{\mathbb{A}} W(\vec{r},\vec{r}')u(\vec{r}')d\vec{r}', \tag{4.1}$$

in which $\mathbb{A}$ may represent for example the surface of a three-dimension diffracting object or the cross-section boundary of a two-dimension (cylindrical) object, v and u are continuous functions in $\mathbb{A}$. The kernel W of the equation is also continuous in $\mathbb{A}$, and $c = 0$ or 1 according to whether the equation is of the first or second kind. The mathematical problem is to determine u if v and W are known [1]. This kind of method is widely employed in many domains of physics [1,2], where it is usual to extend it to cases in which u, v, and W are only peasewise continuous or can even be singular. A typical example of use in electromagnetism can be found by applying the second theorem of Green for expressing the field in a given volume in terms of its values and of the values of its normal derivative on the surrounding surface. In that case, the kernels of the integral equations include combinations of Green's functions and of their normal derivatives on the boundaries of the scattering objects.

The theory of integral equations can be described in a rigorous, elegant and concise way using distribution theory [3-5] that extends derivatives and other differential operators to discontinuous functions (a famous example is the so-called Dirac function that, for the mathematician, should not be called function). The interested reader can find a detailed presentation of rigorous use of the distribution theory in the electromagnetic theory of gratings in [6,7].

The first application of the integral method in grating theory was proposed almost simultaneously for the case of perfectly conducting gratings by Petit and Cadilhac [8], Wirgin [9], and Uretski [10]. The first numerical implementation was reported by Petit [11,12] for TE polarization (called also P, or $E_{//}$, or s polarization). The first extension of this integral equation to the other polarization (TM or S or p) led to a non-integrable kernel. This problem was solved by Pavageau, who proposed for both polarizations new integral equations having continuous and bounded kernels [13].

Soon after, Wirgin [14], Neureuther and Zaki [15], and Van den Berg [16] gave formulations of the integral method applied to gratings made of metals with finite conductivity or dielectrics. The approach was based on the resolution of two coupled integral equations containing two unknown functions. Problems of limited memory storage and time-computation on computers in the late '60s restricted the numerical implementation of this theory to dielectric gratings only, for which very rare numerical results were published. This

restriction was not considered as dramatic by grating specialists at that time. Indeed, it was generally considered that in the visible and infrared regions in which the reflectivity of usual metals exceeds 90%, the model of perfectly conducting grating is accurate. However, development of space optics and astronomy at that time required a precise treatement of the problem of metallic gratings used in the ultraviolet, a domain where the metal reflectivity starts to drop down as approaching electron plasma frequencies. This need required a new approach proposed by Maystre in 1972 [17] by using a single integral equation for a single unknown function. Solving the difficulties in the summation of the series in the kernels and in their integration, the integral method in this formulation was the first one to result in a computer code that was able to correctly model diffraction gratings behaviour over the entire spectrum for almost all commercial gratings profiles [18]. One of the most important conclusions for the practical applications and grating manufacturers was the definite demonstration of the inadequacy of the model of perfectly conducting grating in the near-infrared, visible and ultraviolet regions [19].

However, the method was unable to treat some kinds of gratings, for example gratings having large periods and steep facets (echelle gratings, for example) or gratings with profile that cannot be represented by Fourier series (rod gratings, cavity gratings, etc.). A further development of this approach was proposed by Maystre [20]. It was numerically implemented in the early '90s in the code 'Grating 2000' by the author, which is used in many academic and industrial centers in the world. Other development was required for other exotic cases that started to find applications and thus required theoretical support for modeling. This development covered conical mountings and specially gratings with dielectric multilayer coatings [21,22], buried gratings and bimetallic gratings [23-26].

It must be emphasized that in this chapter, the authors use, without complete demonstrations, some analytic properties of gratings demonstrated in chapter 2. Thus, it is recommended to read this chapter before the present one.

First, we will deal with the most frequent problem: the bare metallic or dielectric grating. Then, extensions will be given to other kinds of gratings like perfectly conducting gratings, dielectric coated gratings or gratings in conical diffraction.

4.2. The integral method applied to a bare, metallic or dielectric grating.

4.2.1. The physical model

The grating surface S of period d separates a region V^+ with real relative electric permittivity and magnetic permeability ε^+ and μ^+ respectively and a region V^- with real or complex relative electric permittivity and magnetic permeability ε^- and μ^- (figure 4.1). The indices n^+ and n^- of these media are given by $n^+ = \sqrt{\varepsilon^+\mu^+}$ and $n^- = \sqrt{\varepsilon^-\mu^-}$. We consider the classical diffraction case with incident wavevector $\vec{k}^i$ lying in the xz plane, i.e. the plane perpendicular to the grooves. The incidence angle θ^i is measured in the counterclockwise sense from the z axis and $\lambda = \frac{2\pi}{k}$ denotes the wavelength of light in vacuum. The ordinate of the top of the profile is denoted by z_0 and unit normal $\vec{N}_S$ is oriented towards V^+. We denote by s the curvilinear abscissa on S, with origin being located at the origin of the Cartesian coordinates, and s_d denoting the curvilinear abscissa of the point of S of abscissa $x = d$.

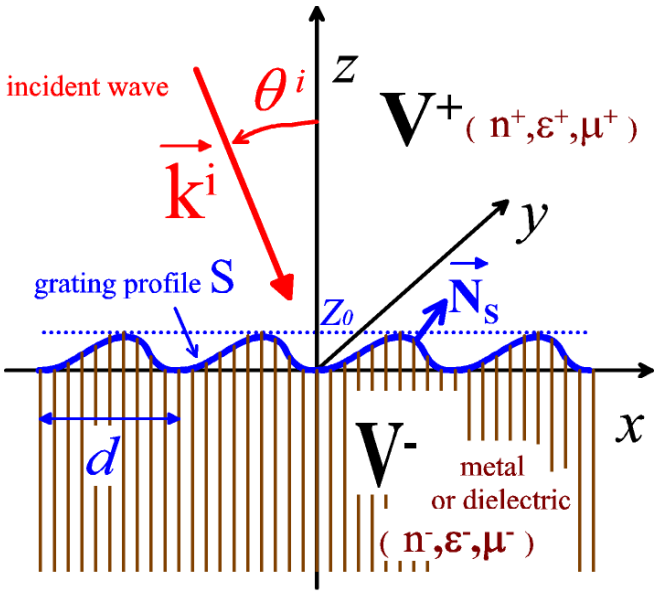


Figure 4.1. Notations

In this chapter, we use the complex notation with a time-dependence in $\exp(-i\omega t)$. Let F denote the y-component of the electromagnetic field. In TE polarization, it stays for the y-component E_y of the electric field, and in TM case for the y-component H_y of the magnetic field.

4.2.2. The boundary value problem.

It is shown in chapter 2 that in that case of classical diffraction, the total field $F^T = \begin{cases} F^{T+} \text{ in } V^+, \\ F^{T-} \text{ in } V^- \end{cases}$ is invariant along the y axis and that it is pseudo-periodic in x:

$$F^T(x+d,z) = F^T(x,z)\exp(i\alpha_0 d), \tag{4.2}$$

with:

$$\alpha_0 = k\, n^+ \sin\theta^i, \qquad k = \frac{2\pi}{\lambda}. \tag{4.3}$$

Moreover, the scattered field is defined by:

$$F = \begin{cases} F^+ = F^{T+} - F^i \quad \text{in } V^+, \\ F^- = F^{T-} \quad \text{in } V^-, \end{cases} \tag{4.4}$$

with the incident field F^i in V^+ being given by:

$$F^i(x,z) = e^{i\alpha_0 x - ikn^+ \cos\theta^i z}. \tag{4.5}$$

The interest of the notion of scattered field is that it satisfies a radiation condition (also called Sommerfeld condition, or outgoing wave condition, see chapter 2) for $z \to \pm\infty$. The radiation

condition states that the scattered field at infinity must remain bounded and must propagate upward in V^+ and dawnward in V^-. The scattered field also satisfies Helmholtz equations:

$$\nabla^2 F^{\pm} + k^2 \left(n^{\pm}\right)^2 F^{\pm} = 0 \quad \text{in } V^{\pm}. \tag{4.6}$$

The invariance along the y axis allows one to reduce the scattering problem to a two-dimension problem while the pseudo-periodicity restricts the study of the field to a single period of the grating. Consequently, it can be considered that $V^{\pm}$ are no more volumes but surfaces extending on a single period of the grating.

In order to periodize the scattered field, we introduce the function $U=\begin{cases} U^+ \text{ in } V^+, \\ U^- \text{ in } V^-, \end{cases}$:

$$U(x,z) = e^{-i\alpha_0 x} F(x,z). \tag{4.7}$$

We denote by $\psi^{\pm}(s)$ the limit values of $U^{\pm}$ on S and $\phi^{\pm}(s)$ the values of $e^{-i\alpha_0 x}\dfrac{dF^{\pm}}{dN_S}$, with $\dfrac{dF^{\pm}}{dN_S}$ being the normal derivative of $F^{\pm}$ on S.

The Helmholtz equations and the radiation condition are not sufficient to define the boundary value problem satisfied by the scattered field F. A third kind of condition must be added: the boundary conditions of the electromagnetic field components across S. The tangential components of the electric and magnetic fields are continuous across this interface, as far as the permittivities and permeabilities of the two media take finite values. For both polarizations, this property yields:

$$\psi^+(s) + \psi^{i}(s) = \psi^-(s), \tag{4.8}$$

with ψ_i being the value of the periodized incident field, obtained from equation (4.5):

$$\psi^i = F^i\left[x(s), z(s)\right]\exp(-i\alpha_0 x(s)) = \exp\left[-ikn^+ \cos\theta^i z(s)\right]. \tag{4.9}$$

Using Maxwell equations, the continuity of the tangential component of the magnetic field (for TE polarization) and that of the electric field (for TM polarization) leads to the following relation:

$$q^+\left[\phi^+(s) + \phi^i(s)\right] = q^-\phi^-(s), \tag{4.10}$$

with

$$\begin{aligned} \phi^i &= \exp(-i\alpha_0 x)\frac{\partial F^i}{\partial N_S} = \\ &= -ikn^+\left(\frac{dz(s)}{ds}\sin\theta^i + \frac{dx(s)}{ds}\cos\theta^i\right)\exp\left(-ikn^+\cos\theta^i z(s)\right), \end{aligned} \tag{4.11}$$

and

$$q^{\pm} = \begin{cases} \dfrac{1}{\mu^{\pm}}, & \text{for TE polarization,} \\ \dfrac{1}{\varepsilon^{\pm}}, & \text{for TM polarization.} \end{cases} \tag{4.12}$$

4.2.3. Integral equation

The theoretical basis of the integral method lies on a general property of the electromagnetic field: the field inside a given surface of the xz plane can be expressed from the values of the field and of its normal derivatives on the curve surrounding the surface, according to the second Green's theorem. The value of $U^{\pm}$ at a point of $V^{\pm}$ of coordinates x and z be deduced from its values on S using equation (4.139) of appendix 4.A:

$$U^{\pm}(x,z) = \pm \int_{s'=0}^{s_d} \left[\mathcal{G}^{\pm}(x,z,s')\phi^{\pm}(s') + \mathcal{N}^{\pm}(x,z,s')\psi^{\pm}(s') \right] ds', \tag{4.13}$$

with

$$\mathcal{G}^{\pm}(x,z,s') = \frac{1}{2id} \sum_{m=-\infty}^{\infty} \frac{1}{\gamma_m^{\pm}} \exp\left\{ imK\left[x - x'(s')\right] + i\gamma_m^{\pm}\left|z - z'(s')\right| \right\}, \tag{4.14}$$

$$\begin{aligned} \mathcal{N}^{\pm}(x,z,s') = \frac{1}{2d} \sum_{m=-\infty}^{\infty} & \left\{ \frac{dx(s')}{ds'} \operatorname{sgn}\left[z - z'(s')\right] - \frac{\alpha_m}{\gamma_m^{\pm}} \frac{dz'(s')}{ds'} \right\} \times \\ & \times \exp\left\{ imK\left[x - x'(s')\right] + i\gamma_m^{\pm}\left|z - z'(s')\right| \right\}, \end{aligned} \tag{4.15}$$

where:

$$\alpha_m = \alpha_0 + mK, \tag{4.16}$$

$$K = \frac{2\pi}{d}, \tag{4.17}$$

$$\gamma_m^{\pm} = \sqrt{\left(kn^{\pm}\right)^2 - \alpha_m^2}, \tag{4.18}$$

with s′ being the curvilinear abscissa on a point of S with coordinates $x'(s')$ and $z'(s')$.

According to section 4.A.4, the values of $\psi^{\pm}(s')$ and $\phi^{\pm}(s')$ are linked by a relation of compatibility. Using eqs. (4.146), (4.147) and (4.148) we obtain:

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^{+}(s,s')\tilde{\phi}^{+}(s') + \mathcal{N}^{+}(s,s')\tilde{\psi}^{+}(s') \right] ds' - \frac{\tilde{\psi}^{+}(s')}{2} = 0, \tag{4.19}$$

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^{-}(s,s')\tilde{\phi}^{-}(s') + \mathcal{N}^{-}(s,s')\tilde{\psi}^{-}(s') \right] ds' + \frac{\tilde{\psi}^{-}(s')}{2} = 0, \tag{4.20}$$

with:

$$\mathcal{G}^{\pm}(s,s') = \frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_m^{\pm}}\exp\left[imK\left(x(s)-x'(s')\right)+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|\right], \tag{4.21}$$

$$\begin{aligned}\mathcal{N}^{\pm}(s,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}&\left[\frac{dx'}{ds'}\operatorname{sgn}\left(z(s)-z'(s')\right)-\frac{\alpha_m}{\gamma_m^{\pm}}\frac{dz'}{ds'}\right]\times\\ \times\exp&\left[imK\left(x(s)-x'(s')\right)+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|\right].\end{aligned} \tag{4.22}$$

Introducing in eq. (4.20) the values of $\psi^-(s)=\psi^+(s)+\psi_i(s)$ and $\phi^-(s)=\frac{q^+}{q^-}\left[\phi^+(s)+\phi_i(s)\right]$ given by the continuity conditions on the grating profile (eqs (4.8), (4.9), (4.10), (4.11) and (4.12)) yields a second integral equations with two unknown functions ψ^+ and ϕ^+:

$$\begin{aligned}\int_{s'=0}^{s_d}&\left\{\frac{q^+}{q^-}\mathcal{G}^-(s,s')\left[\phi^+(s')+\phi^i(s')\right]+\mathcal{N}^-(s,s')\left[\psi^+(s')+\psi^i(s')\right]\right\}ds'\\ &+\frac{\psi^+(s')+\psi^i(s')}{2}=0.\end{aligned} \tag{4.23}$$

Eqs (4.19) and (4.23) constitute a system of two integral equations with two unknown functions, which can be solved on a computer. The amplitudes r_m and t_m of the plane waves reflected and transmitted by the grating can be deduced from the solution of the integral equation using eqs. (4.184) and (4.185) of appendix 4.A:

$$r_m = \frac{1}{2d}\int_{s=0}^{s_d} e^{-imKx(s)-i\gamma_m^+ z(s)}\left[\frac{-i\phi^+(s)}{\gamma_m^+}+\left(\frac{dx(s)}{ds}-\frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\psi^+(s)\right]ds, \tag{4.24}$$

$$t_m = \frac{1}{2d}\int_{s=0}^{s_d} e^{-imKx(s)+i\gamma_m^- z(s)}\left[\frac{i\phi^-(s)}{\gamma_m^-}+\left(\frac{dx(s)}{ds}+\frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\psi^-(s)\right]ds, \tag{4.25}$$

with z_0 being the ordinate of the top of the grating profile. For non-evenescent reflected orders, diffraction efficiencies ρ_m can be obtained using eq. (4.187):

$$\rho_m = \frac{\gamma_m^+}{\gamma_0^+}\left|r_m\right|^2. \tag{4.26}$$

For gratings made of a lossless dielectric material in V^-, transmitted efficiencies can be defined as well:

$$\tau_m = \frac{q^-}{q^+}\frac{\gamma_m^+}{\gamma_0^+}\left|t_m\right|^2. \tag{4.27}$$

In that case the energy balance (see chapter 2) can be expressed by:

$$\sum_{m\in P^+}\rho_m+\sum_{m\in P^-}\tau_m = 1, \tag{4.28}$$

with P^+ and P^- denoting respectively the set of non-evanescent reflected and transmitted orders. The numerical implementation of the integral equations will be described in section 6. In contrast with the two coupled equations obtained in this section, the integral equation obtained by Maystre is unique. This feature requires the definition of a single and well adapted unknown function. The mathematical definition of this function needs the use of tools of applied mathematics described in appendix A. Appendix B contains a mathematical description of this mathematical function and of the integral equation. Here, we give a heuristic description of this function for TE polarization. First, we replace the material in V^- by the same material as in V^+, the entire space being thus homogeneous. It can be shown that it exists one (and only one) distribution of surface current density Φ parallel to the y axis, placed on S (this surface separates now two identical media), which generates in V^+ a field equal to the actual diffracted field in the physical problem. Intuitively, it is easy, from this surface current density, to express in an integral form the actual scattered field in V^+, this current distribution being nothing else than a set of current lines placed in a homogeneous medium. From the expression of the scattered field in V^+, simple mathematical calculations allow one to deduce the scattered field and its normal derivative above S, thus $\psi^+(s)$ and $\phi^+(s)$ from the unique unknown function Φ.

Now, we abandon the field generated by the fictitious surface current density Φ, except the integral expressions of $\psi^+(s)$ and $\phi^+(s)$ containing Φ, and we come back to the actual physical grating problem. The continuity conditions for the tangential components of the field permit the calculation of $\psi^-(s)$ and $\phi^-(s)$ thus, using the second Green theorem, of the actual physical field below S. At that point it has been shown that the four unknown functions contained in the classical theory previously described in this section can be derived from a single one and that, in some way, there is a redundancy in the use of multiple unknown functions. It is easy to understand that this single unknown function can be calculated from a single integral equation. This equation can be obtained for example by writing the continuity on S of the integral expressions of the field in V^+ and V^-.

This method was the first one to show that, in contrast with the second Green theorem, it exists a formula that allows one to express the field inside a given domain from a single function defined on its boundary. This function is neither the field nor its normal derivative, but both can be deduced from it through simple integrals. These integrals automatically satisfy the compatibility condition.

4.3. The bare, perfectly conducting grating

Perfectly conducting gratings were historically the first gratings to be modeled using rigorous electromagnetic theories. They represent accurate models for metallic gratings working in far infrared and microwaves regions. The pioneering works appear in the ‘60s [8] and were followed by many papers. Various formulations of the integral method have been published. They differ either in the form of the integral equation ot in the numerical implementation. A review of this matter may be found in [27, 6, 28].

4.3.1. Perfectly conducting gratings in TE polarization

Two different approaches will be presented in this section. The first one, published by R. Petit and M. Cadilhac in [8], leads to a Fredholm integral equation of the fist kind with a singular

kernel. A version leading to a Fredholm integral equation of the second kind with a non-singular kernel was proposed by Pavageau et al. [13] using the ideas of Maue [29].

The total field in V^+ is pseudo-periodic, it satisfies the Helmholtz equation:

$$\nabla^2 F^{T+} + k^2 \left(n^+\right)^2 F^{T+} = 0 \quad \text{in } V^+, \tag{4.29}$$

and a radiation condition at infinity. Thus we can apply the generalized compatibility condition of section 4.A.5:

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^+(s,s')\Phi^+(s') + \mathcal{N}^+(s,s')\Psi^+(s') \right] ds' + \psi^i(s) = \frac{\Psi^+(s)}{2}, \tag{4.30}$$

with $\Psi^+(s')$ and $\Phi^+(s')$ denoting the limit of the total field on S and its normal derivative. The boundary condition on S is straightforward: the total electric field, which is parallel to the y axis thus tangential to the metal, vanishes on S. This property entails that $\Psi^+(s) = 0$ and thus, eq. (4.30) becomes, in operator notation:

$$\mathcal{G}^+ \Phi^+ = -\psi^i. \tag{4.31}$$

with $\mathcal{G}^+(s,s')$ given by eq. (4.147)

$$\mathcal{G}^+(s,s') = \frac{1}{2id} \sum_{m=-\infty}^{\infty} \frac{1}{\gamma_m^+} \exp\left[imK\left(x(s) - x'(s')\right) + i\gamma_m^+ \left| z(s) - z'(s') \right| \right]. \tag{4.32}$$

This is a Fredholm integral equation of the first kind, with a singular kernel.

The amplitudes of the reflected waves are deduced from eq. (4.186):

$$r_m = \frac{1}{2id\gamma_m^+} \int_{s=0}^{s_d} \exp\left[-imKx(s) - i\gamma_m^+ z(s) \right] \Phi^+(s)\, ds. \tag{4.33}$$

The efficiencies $\rho_m = \frac{\gamma_m^+}{\gamma_0^+} |r_m|^2$ satisfy the energy balance relation:

$$\sum_{P^+} \rho_m = 1, \tag{4.34}$$

with P^+ denoting the set of non-evanescent orders.

An integral equation of the second kind with a regular continuous kernel can be found using the same function Φ^+. It is shown in section 4.A5 that the normal derivative $\frac{dF^+}{dN_S}$ can be calculated in that case (eq. (4.172)). This integral equation can be written either by writing that $\frac{dF^+}{dN_S} = \phi^+ e^{i\alpha_0 x}$ is equal to $\left(\Phi^+ - \phi^i\right) e^{i\alpha_0 x}$. The final equation is given by:

$$\frac{\Phi^+(s)}{2} = \phi^i(s) + \int_{s'=0}^{s_d} \mathcal{K}(s,s') e^{i\alpha_0 x} \Phi^+(s)\, ds', \tag{4.35}$$

with:

$$\mathcal{K}(s,s') = \frac{1}{2d}\sum_{m=-\infty,+\infty}\left[\operatorname{sgn}(z-z')\frac{dx}{ds} - \frac{\alpha_m}{\gamma_m^+}\frac{dz}{ds}\right] e^{imK(x-x')+i\gamma_m^+|z-z'|}. \tag{4.36}$$

4.3.2. Perfectly conducting gratings for TM polarization

Once again, the generalized compatibility condition is used (eq. (4.158):

$$\int_{s'=0}^{s_d}\left[\mathcal{G}^+(s,s')\Phi^+(s') + \mathcal{N}^+(s,s')\Psi^+(s')\right]ds' + \psi^i(s) = \frac{\Psi^+(s)}{2}. \tag{4.37}$$

In that case too, the tangential component of the electric field vanishes on the profile. It is shown in chapter 2 from Maxwell equations that this condition entails that the normal derivative of the total field vanishes on S, thus :

$$\Phi^+ = 0, \tag{4.38}$$

so that

$$\int_{s'=0}^{s_d}\mathcal{N}^+(s,s')\Psi^+(s')ds' + \psi^i(s) = \frac{\Psi^+(s)}{2}, \tag{4.39}$$

with $\mathcal{N}^+(s,s')$ given by eq. (4.22) and (4.148):

$$\begin{aligned}\mathcal{N}^+(s,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}&\left[\frac{dx'}{ds'}\operatorname{sgn}(z(s)-z'(s')) - \frac{\alpha_m}{\gamma_m^\pm}\frac{dz'}{ds'}\right]\times \\ \times\exp&\left[imK(x(s)-x'(s')) + i\gamma_m^+|z(s)-z'(s')|\right].\end{aligned} \tag{4.40}$$

The Fredholm integral equation of the second kind with a regular continuous kernel is very close to that obtained for TE polarization (eq. (4.35)).

The amplitudes of the reflected waves are deduced from eq. (4.186):

$$r_m = \frac{1}{2d}\int_{s=0}^{s_d}\exp\left[-imKx(s) - i\gamma_m^+ z(s)\right]\left(\frac{dx(s)}{ds} - \frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\Psi(s)ds, \tag{4.41}$$

As for TE polarization, the efficiencies $\rho_m = \frac{\gamma_m^+}{\gamma_0^+}|r_m|^2$ satisfy the energy balance relation:

$$\sum_{P^+}\rho_m = 1. \tag{4.42}$$

4.4. Multiprofile gratings

The use of dielectric coatings has many applications even for diffraction gratings use. For example, metallic gratings are covered by a thin layer of dielectric material in order to avoid oxidation of the metal. Dielectric gratings can require an antireflection coating consisting of a thin layer or a stack of layers. Conversely, a stack of layers is used to increase the metal

reflectivity or even to replace it, in order to reduce the absorption of light beams for high power laser applications.

Up to our knowledge, the first numerical results in the study of coated gratings was made by Van de Berg [16] for a single-layer perfectly conducting grating with the layer filling up the space between the grating surface and a plane surface. Although at that time the interest in such geometry remained mostly academic, further development of technology made it possible to fabricate such layer by dielectric coating and polishing. Another important application comes from the process of replication of dielectric gratings using an epoxy layer to transfer the replica to a plane surface of the substrate, or to have epoxy as grating layer itself.

After this initial work, two integral methods were proposed. The first one [20] is theoretically able to deal with an arbitrary multilayer grating without limitations concerning the shape of the profile or the conductivity of the layers. The second method [21] can deal with a multilayer grating without interpenetration of the profiles.Botten has solved the problem with a single-profile grating that has a stack of plane layers below and, eventually, above it [22, 23], by introducing a new form of Green's function, adapted to the multilayer system, which leeds to a single integral equation.

In what follows, we will at first describe the method for a single interface inside a stack, when the layer is relatively thin so that the upper and the lower interface interpenetrate. It is important to distinguish the two cases, with and without interpenetration, because in the latter case, it is possible to define a plane layer in between that does not cross the upper or the lower interface. This possibility enables one to use the plane-wave Rayleigh expansion of the electromagnetic field between the interfaces, whereas in the former case it is necessary to write and to solve a system of coupled integral equations that link the field components on the top and bottom of each layer.

4.4.1. Thin-layer gratings

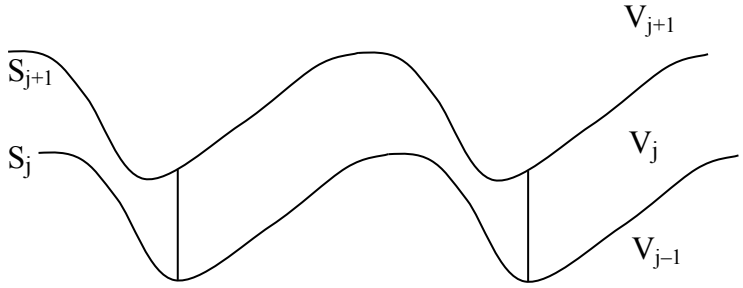


Figure 4.2. Single layer inside a stack of a multilayer grating

Let us consider the case of a multilayer grating having profiles that interpenetrate. In other words, it is impossible to introduce inside the layer a plane surface that does not cross one of the profiles. Then it is impssible to use the plane-wave expansion between the interfaces and we are led to solve integral equations that are coupled on the two interfaces of each layer.

We introduce in figure 4.2 a grating made of M materials (numbered from 0 to M, separated by M-1 profiles (numbered from 1 to M). We introduce the following functions:

$$U_j = \begin{cases} F_j e^{-i\alpha_0 x} - F^i e^{-i\alpha_0 x}\delta_{j,M} & \text{in } V_j \\ 0 & \text{elsewhere.} \end{cases} . \tag{4.43}$$

The function U_j inside each layer can be expressed from the fields and normal derivatives on the lower and upper interface, for j=1, M-1:

$$\begin{aligned} U_j(x,y) = &\int_{S_j} \mathcal{G}_j^+(x,y,s_j')\phi_j^+(s_j')ds_j' + \int_{S_j} \mathcal{N}_j^+(x,y,s_j')\psi_j^+(s_j')ds_j' \\ &- \int_{S_{j+1}} \mathcal{G}_{j+1}^-(x,y,s_{j+1}')\phi_{j+1}^-(s_{j+1}')ds_{j+1}' - \int_{S_{j+1}} \mathcal{N}_{j+1}^-(x,y,s_{j+1}')\psi_{j+1}^-(s_{j+1}')ds_{j+1}'. \end{aligned} \tag{4.44}$$

The expression being limited to the second one if j = 0 and to the first one for j = M.

The functions derived from the Green functions in the various materials depend on the interface number:

$$\mathcal{G}_j^{\pm}(x,y,s') = \frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_{j,m}^{\pm}}\exp\left\{imK\left[x - x'\left(s_j'\right)\right] + i\gamma_{j,m}^{\pm}\left|z - z'\left(s_j'\right)\right|\right\}, \tag{4.45}$$

$$\begin{aligned} \mathcal{N}_j^{\pm}(x,y,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}&\left\{\frac{dx'(s')}{ds'}\operatorname{sgn}\left[z - z'\left(s_j'\right)\right] - \frac{\alpha_m}{\gamma_{j,m}^{\pm}}\frac{dz'(s_j')}{ds'}\right\}\times \\ &\times\exp\left\{imK\left[x - x'\left(s_j'\right)\right] + i\gamma_{j,m}^{\pm}\left|z - z'\left(s_j'\right)\right|\right\}, \end{aligned} \tag{4.46}$$

with:

$$\gamma_{j,m}^{+} = \gamma_{j+1,m}^{-} = \sqrt{\left(kn_j\right)^2 - \alpha_m^2}\,. \tag{4.47}$$

It can be shown it the same manner as in eq. (4.146) that a compatibility condition on the j^{th} interfacewritten in a matrix form is given by:

$$\frac{\psi_j^+}{2} = \mathcal{G}_j^+\phi_j^+ + \mathcal{N}_j^+\psi_j^+ - \mathcal{G}_{j,j+1}^-\phi_{j+1}^- - \mathcal{N}_{j,j+1}^-\psi_{j+1}^-\,. \tag{4.48}$$

Another compatibility equation is obtained on the $(j+1)^{th}$ interface:

$$\frac{\psi_{j+1}^-}{2} = \mathcal{G}_{j+1,j}^+\phi_j^+ + \mathcal{N}_{j+1,j}^-\psi_j^+ - \mathcal{G}_{j+1}^-\phi_{j+1}^- - \mathcal{N}_{j+1}^-\psi_{j+1}^-\,. \tag{4.49}$$

In eqs. (4.48) and (4.49), we use the double-index Greens functions that are derived on two consecutive profiles: $\mathcal{G}_{j,j+1}^- \equiv \mathcal{G}_j^-(s_j, s_{j+1}')$, $\mathcal{G}_{j+1,j}^+ \equiv \mathcal{G}_j^+(s_{j+1}, s_j')$, and similarly for $\mathcal{N}$.

The computability equation becomes, for the upper and lower media:

$$\frac{\psi_M^+}{2} = \mathcal{G}_M^+\phi_M^+ + \mathcal{N}_M^+\psi_M^+\,, \tag{4.50}$$

$$\frac{\psi_1^-}{2} = -\mathcal{G}_1^-\phi_1^- - \mathcal{N}_1^-\psi_1^-\,. \tag{4.51}$$

By combining eqs (4.48) with eq. (4.49), we obtain the link between the unknows on the upper and on the lower interface of the j^{th} layer, for j=1, M-1:

$$\begin{pmatrix} \mathcal{N}^-_{j,j+1} & \mathcal{G}^-_{j,j+1} \\ \mathcal{N}^-_{j+1}+\frac{\mathbb{I}}{2} & \mathcal{G}^-_{j+1} \end{pmatrix}\begin{pmatrix} \psi^-_{j+1} \\ \phi^-_{j+1} \end{pmatrix} = \begin{pmatrix} \mathcal{N}^+_{j}-\frac{\mathbb{I}}{2} & \mathcal{G}^+_{j} \\ \mathcal{N}^+_{j+1,j} & \mathcal{G}^+_{j+1,j} \end{pmatrix}\begin{pmatrix} \psi^+_{j} \\ \phi^+_{j} \end{pmatrix}. \tag{4.52}$$

This equation gives the transmission operator of the unknown amplitudes across the j^{th} layer, i.e. from the j^{th} to the $(j+1)^{st}$ interface, for j=1,M-1:

$$T_{j+1,j} = \begin{pmatrix} \mathcal{N}^-_{j,j+1} & \mathcal{G}^-_{j,j+1} \\ \mathcal{N}^-_{j+1}+\frac{\mathbb{I}}{2} & \mathcal{G}^-_{j+1} \end{pmatrix}^{-1}\begin{pmatrix} \mathcal{N}^+_{j}-\frac{\mathbb{I}}{2} & \mathcal{G}^+_{j} \\ \mathcal{N}^+_{j+1,j} & \mathcal{G}^+_{j+1,j} \end{pmatrix}. \tag{4.53}$$

The transmission operator includes an inverse operator. Numerically, this inversion leads to the inversion of a matrix, as we will see in section 4.6.

The transmission matrix across the j^{th} interface for j = 1, …, M–1 is obtained using the continuity of the tangential and normal field components, as given by eqs.(4.8) and (4.10):

$$\begin{pmatrix} \psi^+_j \\ \phi^+_j \end{pmatrix} = T^{+-}_j \begin{pmatrix} \psi^-_j \\ \phi^-_j \end{pmatrix}, \qquad T^{+-}_j = \begin{pmatrix} \mathbb{I} & 0 \\ 0 & \frac{q^-_j}{q^+_j}\mathbb{I} \end{pmatrix}, \; j \neq M. \tag{4.54}$$

The advantage of this presentation is that there are no exponentially growing terms in the transmission matrices, since all the components of the two variable functions contain only scattered propagating or decreasing evanescent waves, in contrast with the other methods (differential, Fourier modal, Rayleigh, etc.). However, this formulation requires calculating the cross-layer functions between the interfaces, which leads to computation times equal to those of single-interface functions. As a consequence, the total computation time is almost multiplied by a factor 2 with respect to the case where cross-layer kernels can be avoided, as discussed in the next section.

Finally, it can be deduced from eqs. (4.53) and (4.54):

$$\begin{pmatrix} \psi^+_M \\ \phi^+_M \end{pmatrix} + \begin{pmatrix} \psi^i \\ \phi^i \end{pmatrix} = T^{+-}_M T_{M,M-1} \quad \ldots \quad T^{+-}_2 T_{2,1} T^{+-}_1 \begin{pmatrix} \psi^-_1 \\ \phi^-_1 \end{pmatrix}. \tag{4.55}$$

Finally, eqs. (4.55), (4.50) and (4.51) form an operator system of 4 equations with 4 unknown functions, which can be solved on a computer after representing each operator by a matrix, as described in section 6.

4.4.2. Profiles without interpenetration

This case is simpler than the situation in sec.4.1, because it is possible to use the plane-wave expansion between the grating profiles and thus to decouple the integral presentation used in eq.(4.44). The idea is illustrated in figure 4.3.

If it is possible to introduce a plane layer between the profiles, the plane wave expansion is valid inside this layer. The advantage is that the plane waves (propagating and evanescent) that represent each diffraction order m can be easily separated into to sets: (i) upgoing waves having amplitudes of the y-component of the field equal to $r_{j,m}$ that are

generated by the lower surface S_j, (ii) downgoing waves with amplitudes $t_{j,m}$ generated by the upper grating surface S_{j+1}.

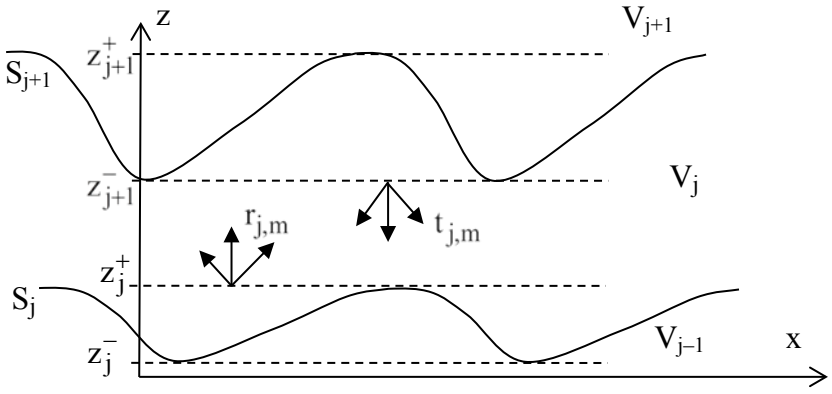


Figure.4.3. Layer with two profiles that are separable by a plane layer

Let us first rewrite eqs.(4.184) at $z > z^+_j$:

$$r_{j,m} = \int_{s=0}^{s_{j,d}} \left[N^+_{j,m}\left(s_j\right)\psi^+_j\left(s_j\right) + G^+_{j,m}\left(s_j\right)\phi^+_j\left(s_j\right)\right] ds_j \,, \tag{4.56}$$

with

$$G^+_{j,m}\left(s_j\right) = \frac{1}{2id\gamma^+_{j,m}} \exp\left[-imKx\left(s_j\right) - i\gamma^+_{j,m} z\left(s_j\right)\right], \tag{4.57}$$

$$N^+_{j,m}\left(s_j\right) = \frac{1}{2d}\left[\frac{dx\left(s_j\right)}{ds_j} - \frac{\alpha_m}{\gamma^+_{j,m}} \frac{dz\left(s_j\right)}{ds_j} \right] \exp\left[-imKx\left(s_j\right) - i\gamma^+_{j,m} z\left(s_j\right)\right].$$

and with s_j denoting the curvilinear abscissa on the j^{th} profile, $s_{j,d}$ being the curvilinear abscissa of the point of S_j of abscissa x = d.

We then can directly use eq.(4.48) to express the field on the interface $z = z(s_{j+1})$:

$$\frac{\psi^-_{j+1}\left(s_{j+1}\right)}{2} = \sum_m \exp\left[imKx(s_{j+1}) + i\gamma^+_{j,m} z(s_{j+1})\right] r^+_{j,m} - \mathcal{G}^-_{j+1}\phi^-_{j+1} - \mathcal{N}^-_{j+1}\psi^-_{j+1} \,. \tag{4.58}$$

Let us consider the sum in eq. (4.58):

$$\zeta_{j+1} = \sum_m \exp\left[imKx(s_{j+1}) + i\gamma^+_{j,m} z(s_{j+1})\right] r_{j,m} \,. \tag{4.59}$$

Inserting in this equation the value of $r_{j,m}$ given by eq. (4.56) yields:

$$\zeta_{j+1}(s_{j+1}) = \sum_m \exp\left[imKx(s_{j+1}) + i\gamma^+_{j,m} z(s_{j+1})\right] \int_{s_j=0}^{s_{j,d}} \left[N^+_{j,m}\left(s_j\right)\psi^+_j\left(s_j\right) + G^+_{j,m}\left(s_j\right)\phi^+_j\left(s_j\right)\right] ds_j, \quad (4.60)$$

which can be written in operator form:

$$\zeta_{j+1} = \mathbf{N}^+_{j,j+1}\psi^+_j + \mathbf{G}^+_{j,j+1}\phi^+_j, \quad (4.61)$$

where the operators $\mathbf{N}^+_{j,j+1}$ and $\mathbf{G}^+_{j,j+1}$ are obtained from a summation in m and an integral in s_j. Thus we can write eq. (4.58) in the operator form:

$$\frac{\psi^-_{j+1}}{2} = \mathbf{N}^+_{j,j+1}\psi^+_j + \mathbf{G}^+_{j,j+1}\phi^+_j - \boldsymbol{\mathcal{G}}^-_{j+1}\phi^-_{j+1} - \boldsymbol{\mathcal{N}}^-_{j+1}\psi^-_{j+1}. \quad (4.62)$$

The second relation comes from eq. (4.49) by using the amplitudes of the diffraction orders coming down from the upper interfaces and valid below z = z^-_{j+1}. Following the same lines as in the previous paragraph yields:

$$\frac{\psi^-_j}{2} = \mathbf{N}^-_{j+1,j}\psi^-_{j+1} + \mathbf{G}^-_{j+1,j}\phi^-_{j+1} + \boldsymbol{\mathcal{G}}^+_j\phi^+_j + \boldsymbol{\mathcal{N}}^+_j\psi^+_j. \quad (4.63)$$

The transmission matrix between the j^{th} and the $j+1^{st}$ interface takes a form similar to the case of interpenetrating layers, eq.(4.53). However, the difference is essential, because each series used in (4.63) is evaluated on a single interface:

$$T_{j+1,j} = \begin{pmatrix} \mathbf{N}^-_{j+1,j} & \mathbf{G}^-_{j+1,j} \\ \boldsymbol{\mathcal{N}}^-_{j+1} + \frac{\mathbb{I}}{2} & \boldsymbol{\mathcal{G}}^-_{j+1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\mathcal{N}}^+_j - \frac{\mathbb{I}}{2} & \boldsymbol{\mathcal{G}}^+_j \\ \mathbf{N}^+_{j,j+1} & \mathbf{G}^+_{j,j+1} \end{pmatrix}. \quad (4.64)$$

The second difference is that the exponential terms are explicitly given in the $\mathbf{N}^+_{j,j+1}$, $\mathbf{N}^-_{j+1,j}$, $\mathbf{G}^+_{j,j+1}$ and $\mathbf{G}^-_{j+1,j}$ through the functions $\exp\left[imKx(s_{j+1}) + i\gamma^+_{j,m} z(s_{j+1})\right]$ and $\exp\left[imKx(s_j) - i\gamma^+_{j,m} z(s_j)\right]$ in such a way that it can be extracted from each of these operators a part containing all the growing and decreasing exponential terms, which allow a much better stability of the numerical implementation through adequate treatments, for example the S-matrix algorithm described in appendix A and B of Chapter 7.

4.5. Gratings in conical mounting

When classical gratings with one-dimensional periodicity are used with incidence plane perpendicular to the grooves, the diffraction orders lie in the same plane. Off-plane incidence brings the diffraction orders out of the plane and their directions lie on a cone , which explains the term of conical diffraction. One of the first experimental works can be found in [24, 25]. The use of conical mount is typical for concave gratings and in some spectrographs aiming to separate off-plane the incident and the diffracted directions.

The first theoretical studies were made in 1971 using the integral [26] and the differential [30] methods. An interesting theoretical development came in 1972 when Maystre and Petit demonstrated that under special conditions, mechanically ruled perfectly conducting gratings can have very high and constant efficiency over a large spectral domain [31]. They also established *the theorem of invariance* [32] that gives an expression of the diffraction efficiency of a perfectly conducting gratings in conical mounting, expressed as a linear combination of efficiencies in an in-plane (classical) mount for the two fundamental polarizations. Since the theorem is not valid for finitely conducting metals, later development of the integral method allowed studies of diffraction gratings behaviour in conical mount when working in the UV and visible [33, 34]. An interested reader can find the demonstration of the invariance theorem in [7, 33].

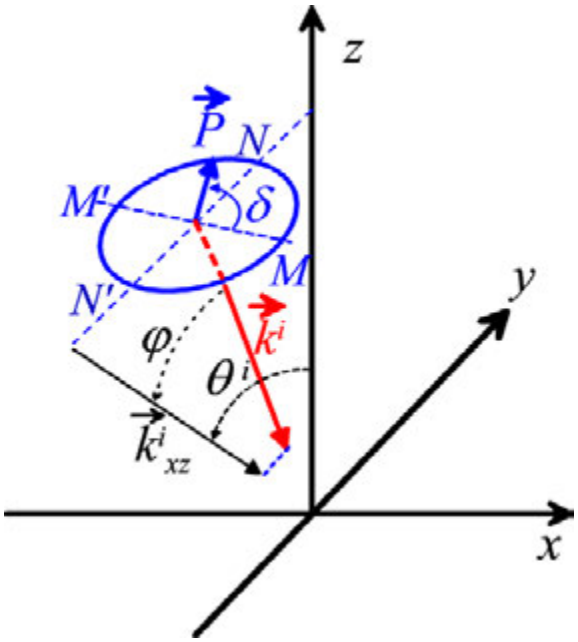


Figure 4.4. Parameters of the incident wave in conical mount.The angle φ *denotes the angle between the incident wavevector* $\vec{k}^i$ *and its projection* $\vec{k}^i_{xz}$ *on the xz plane. The angle* θ^i *is the angle between the z axis and* $\vec{k}^i_{xz}$. *In order to define the polarization of the incident field, we construct the circle MNM'N' in the plane perpendicular to the incident wavevector* $\vec{k}^i$, *with the continuation of NN' intersecting the z axis and MM' being perpendicular to NN'. The polarization angle* δ *is the angle between M'M and the direction of the incident wavevector* $\vec{k}^i$.

The notations are summarized in figure 4.4. The mathematical formulation of the invariance theorem takes the form of an equivalence between the conical case and an associated classical case:

(i) conical case:

With incident angles θ^i and φ, incident polarization angle δ, incident wavelength λ, the efficiencies in the various orders are denoted by $\rho_m\left(\theta^i, \varphi, \delta, \lambda\right)$.

(ii) fictitious equivalent classical case:

The wavelength is increased to $\lambda' = \lambda / \cos\varphi$, the angle φ' is now equal to 0 (in-plane incidence), the angle $\theta'^i = \theta^i$, then efficiencies in TE and TM polarizations are denoted by $\rho_m^{TE}\left(\theta^i, \lambda'\right)$ and $\rho_m^{TM}\left(\theta^i, \lambda'\right)$, respectively. It can be noticed that the projection $\vec{k}^i_{xz}$ of the

wavevector of the incident wave on the xz plane in conical mount identifies with the wavevector of the incident waves in the fictitious equivalent case.

The invariance theorem states that:

$$\rho_m\left(\theta^i,\varphi,\delta,\lambda\right)=\left(\cos\delta\right)^2\rho_m^{TE}\left(\theta^i,\lambda'\right)+\left(\sin\delta\right)^2\rho_m^{TM}\left(\theta^i,\lambda'\right). \tag{4.65}$$

It is to be noticed that, as for the incident wavevector, the projections of the wavevectors of the scattered waves on the xz plane in conical mount are identical to the wavevectors of the scattered waves in the fictitious equivalent case.

4.6. Numerical tools for an efficient numerical implementation

4.6.1. Integration schemes for the integral equation

All the integral equations in this chapter link the value of an unknown function u(s) at a given point of S to its value on its entire domain of definition:

$$c\,u(s)=v(s)+\int_0^{s_d}W(s,s')u(s')ds', \tag{4.66}$$

where all functions are periodical, with v and W being known functions, W being possibly singular but integrable. The constant c takes values 0 or 1 according to whether the integral equation is of the first or second kind.

There are many different ways to solve such an equation for the grating problem. Pavageau et al. proposed an iterative method [13] that does not require any matrix inversion, like the well known Born method for scattering problems. Unfortunately, it may diverge [35, 36].

A well known general methods is based on the periodicity of all functions of the equation, which allows a projection of these functions and of the equation on the Fourier space:

$$u(s)=\sum_m u_m\exp(imK_s s),\quad K_s=\frac{2\pi}{s_d}, \tag{4.67}$$

and similar expressions for v and W. The integral equation is transformed into a linear system of algebraic equations:

$$\sum_m\left(W_{nm}-c\delta_{nm}\right)u_m=v_n,\quad\forall n, \tag{4.68}$$

which can be solved numerically after truncation. However, this approach requires computing a double Fourier decomposition:

$$W_{nm}=\int_0^{s_d}\int_0^{s_d}W(s,s')u(s')\exp(-inK_s s-imK_s s')ds\,ds'. \tag{4.69}$$

The method has been applied to gratings with profiles consisting of few straight segments, because in that case the double Fourier integral can be calculated in closed form. It is so for triangular profiles [11] or trapezoidal gratings [37].

The most widespread method is the so-called point-matching (or discretization) method. Instead ot using discrete Fourier components, the unknown function is discretized on the

grating profile and represented by its vales $u_j = u(s_j)$ inside the interval of integration. Ther integration process leads toan equation quite similar to eq.(4.68)

$$\sum_p \left(W_{jp} - c\delta_{jp} \right) u_p = v_j, \quad \forall j \in [1, P], \tag{4.70}$$

with P being the number of matching points. It is worth noting that the value of W_{jp} may differ from the value of $W\left(s_j, s_p\right)$ obtained through the rectangular rule of integration (multiplied by the weight of integration) and can require much more complicated treatments, specially if W is singular. The sophistication of this treatment is one of the decisive keys for the precision of the solution of the integral solution. The second key is the analytical study of the kernel, which is described in the next two sections.

When W is regular, continuous, with a continuous first derivative, the rectangular rule of integration is quite precise since the function to be integrated is periodic and it can be noticed that the rectangular and trapezoidal rules are completely equivalent in that case. However, the derivative of W(s,s') is generally discontinuous when s = s' and a trapezoidal rule or higher order treatment provide a better precision [38, 39]. In what follows we assume that u(s) is a continuous function. This is obviously the case when the grating profile has no edges. Several more detailed arguments in favour of the rectangular rule can be found in [6, 7]. Let us shortly repeat one of them. We consider an integral of a periodical continuous function a(s):

$$a_0 = \int_0^{s_d} a(s) ds \,. \tag{4.71}$$

The exact integral is equal to the 0^{th} term in the Fourier expansion of a(s):

$$a(s) = \sum_m a_m \exp(imK_s s) \,. \tag{4.72}$$

Simple calculations show that when using the rectangular rule with P discretization points, the integration error is proportional to the P^{th} Fourier coefficient of a(x). Since it is continuous, the Fourier series converges like $1/P^2$ at least.

The implementation of the rectangular rule is very simple. We define the values of W_{jp} in the following manner:

$$W_{j,p} = \frac{s_d}{P} W\left(\frac{j}{P} s_d, \frac{p}{P} s_d \right), \quad j, p = 1, ..., P-1 \,. \tag{4.73}$$

Using this result, the product of the unknown functions u(s) with the non-singular parts of the kernels can be integrated in the form of a simple matrix product:

$$\int_0^{s_d} W(s_j, xs) u(s') ds' \approx \sum_p W_{j,p} u_p \,. \tag{4.74}$$

In the case of multilayer gratings with profiles interpenetration, there are two main situations that can complicate the numerical evaluation of the cross-layer functions which link the fields and normal derivatives on both sides of a layer:

(i) The wavelength λ is much larger than the layer thickness t. In that case the profiles are located too close to each other, compared to λ, so that the functions $\mathcal{G}$ and $\mathcal{N}$ become large in

modulus for $s_{j+1} \to s_j$. This behaviour has the same origin as the singularities of the kernels $\mathcal{G}$ and $\mathcal{N}$ for a bare grating , which will be discussed in section 6.3 and can be eliminated in a similar manner. Numerical results show no problems as long as layer thicknesses exceed λ/20.

(ii) The period d is much greater than the layer thickness t. If the wavelength is not too much larger than t, then the kernels have no singularities, but their moduli present peaks when the distance $\left|P_{j+1,p}P_{j,q}\right|$ between two points located on the two different profiles is small with respect to the discretization distance between the points located on the same profile. The width of these maxima is of the order of magnitude of the layer thickness, thus the correct implementation the trapezoidal integration rule requires the distance between two consecutive points of the profile discretization $\Delta = \left|s_{j,p} - s_{j,p-1}\right|$ to be less than the width of the maxima. As a rule of thumb, if $\Delta \approx d/N_p$, where N_p is the number of integration points, then its lower limit is determined by the relation:

$$N_{p,min} \propto \frac{d}{t}. \tag{4.75}$$

Thus, for echelles, a values of d of about 10 μm and t of 20 nm requires the number of integration points to exceed 500. Note that N_p directly determines the number of unknown values of ϕ_j and ψ_j and thus the size of the matrices to be used. Practically, it is not worth nowadays for N_p to exceed 10 000, because of the computation time, memory requirements, round-off errors and limited digits. As a consequence, it is necessary to find another way of integration instead of the trapezoidal rule.

There is another problem that can come from the matrix inversion in the construction of the transmission matrix between the layers, eq.(4.53). Contrary to the transmission matrix that contains growing and decreasing exponential terms in the plane wave expansion, (thus requires some type of recursive algorithm to contain the contamination of the growing exponentials, S-matrix algorithm, for example, see appendix 4.C), the distance between the profiles in the z-direction that appears in the kernels in eqs. (4.45) and (4.46) restricts the terms to only propagation or evanescently decreasing ones. However, the matrix inversion of these terms that is required in eq.(4.53) can create exponentially growing terms. Two situations can appear:

1. The matrix inversion in eq.(4.53) can be done without numerical problems. This happens when the layer thickness is not quite large. In that case it is possible to progress upwards in the stack of layers by following the S-matrix algorithm.
2. The matrix inversion does not work. This could happen if the distance between two consecutive interface is large. Two different geometries can be concerned:

(i) there is no interpenetration of these two profiles. In that case one can easily apply the technique described in the next section.

(ii) there is interpenetration of two very deep interfaces. It is possible to use directly eq.(4.52) in the S-matrix algorithm without inverting the matrix to calculate the entire T-matrix in eq. (4.53). The formulation of the S-matrix algorithm to an equation having the form given in (4.52) is quite similar to the classical aplication, but it requires one additional iteration step. The advantage is that it avoids the inversion of small terms that can lead to singular matrices. This special technique is given in Appendix 4.C and is not quite popular, but can be used in other methods that apply for multilayer stack, for example, in the coordinate transformation method.

4.6.2. Summation of the kernels

There are several problems in the calculation of the functions included in eqs (4.19) and (4.23):

$$\mathcal{G}^{\pm}(s,s') = \frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_m^{\pm}}e^{imK\left[x(s)-x'(s')\right]+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|}, \tag{4.76}$$

$$\mathcal{N}^{\pm}(s,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}\left[\frac{dx'}{ds'}\operatorname{sgn}\left(z(s)-z'(s')\right)-\frac{\alpha_m}{\gamma_m^{\pm}}\frac{dz'}{ds'}\right]e^{imK\left[x(s)-x'(s')\right]+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|}, \tag{4.77}$$

$$\mathcal{K}(s,s') = \frac{1}{2d}\sum_{m=-\infty,+\infty}\left[\operatorname{sgn}\left(z-z'\right)\frac{dx}{ds}-\frac{\alpha_m}{\gamma_m^{+}}\frac{dz}{ds}\right]e^{imK\left[x(s)-x'(s')\right]+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|}, \tag{4.78}$$

with:

$$\alpha_m = \alpha_0 + mK, \tag{4.79}$$

$$\gamma_m^{\pm} = \sqrt{\left(kn^{\pm}\right)^2-\alpha_m^2}\,. \tag{4.80}$$

For the sake of simplicity, we assume here that $n^{\pm}=1$ and we cancel the superscript $\pm$ in $\gamma_m^{\pm}$, $\mathcal{G}^{\pm}$ and $\mathcal{N}^{\pm}$.

Let us at first evaluate the asymptotic values of γ_m and α_m for large values of m:

$$\begin{aligned}\alpha_m &\underset{m\to\infty}{\to} mK,\\ \gamma_m \approx i\left|\alpha_m\right| - \frac{ik^2}{2\left|\alpha_m\right|} &\underset{m\to\infty}{\to} i\left|\alpha_0+mK\right|,\\ \frac{1}{\gamma_m} \approx \frac{1}{i\left|\alpha_m\right|} + \frac{k^2}{2i\left|\alpha_m\right|^3} &\underset{m\to\infty}{\to} \frac{1}{i\left|m\right|K}.\end{aligned} \tag{4.81}$$

We consider the function

$$\mathcal{G}(s,s') = \frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_m}\exp\left\{imK\left[x(s)-x(s')\right]+i\gamma_m\left|z(s)-z(s')\right|\right\}. \tag{4.82}$$

At point $s=s'$, it is obvious that the sum does not converge since the terms decrease in $1/|m|$. Of course, a very slow convergence can be expected when the two points are close to each other. Different techniques have been proposed to accelerate the convergence. Neureuther and Zaki [15] have employed a transformation technique based on the use of Mellin transforms. Other authors have proposed accelerating processes [40-43]. Here we describe a direct approach [7]. Let us determine at first the asymptotic expression of the kernel. If we replace (4.81) into eq.(4.82), we obtain the asymptotic term in the sum:

$$\begin{aligned}\mathcal{G}_\infty(s,s') = e^{\alpha_0[x(s)-x(s')]} \sum_{m=-1}^{-\infty} \frac{1}{4\pi m} e^{mK\left[x(s)-x(s')+i|z(s)-z(s')|\right]} \\ - e^{-\alpha_0[x(s)-x(s')]} \sum_{m=1}^{\infty} \frac{1}{4\pi m} e^{-mK\left[x(s)-x(s')-i|z(s)-z(s')|\right]},\end{aligned} \tag{4.83}$$

which contains two sums of the form $\sum_{m=1}^{\infty} \xi^m / m$ and can be summed in closed form:

$$\begin{aligned}\mathcal{G}_\infty(s,s') = \frac{1}{4\pi} e^{\alpha_0[x(s)-x(s')]} \log\left(1 - e^{-K\left[x(s)-x(s')+i|z(s)-z(s')|\right]}\right) \\ + \frac{1}{4\pi} e^{-\alpha_0[x(s)-x(s')]} \log\left(1 - e^{-K\left[x(s)-x(s')-i|z(s)-z(s')|\right]}\right).\end{aligned} \tag{4.84}$$

The calculation of the kernel in (4.82) is achieved by subtracting the asymptotic value:

$$\mathcal{G} = \mathcal{G}_\infty + \left(\mathcal{G} - \mathcal{G}_\infty\right). \tag{4.85}$$

The first term in the right-and side is explicitely given in (4.84). It is singular for s = s'e. This singularity is integrable and is be treated in the next section.

The term between parenthesis in eq.(4.85) is obtained by combining the terms in the sums in eqs.(4.82) and (4.83). As far as the second one is the asymptotic value of the former, the series converges, whatever the values of s and s'. Furthermore, it is possible to show that by combining the terms m and –m in the sum, we finally obtain a rapidly converging series, whose terms decrease as m^{-3} when $s = s'$, as shown later in eq.(4.87) Moreover, in this case this series is continuous and its value is simply given by:

$$\left(\mathcal{G} - \mathcal{G}_\infty\right)_{|s=s'} = \frac{1}{2id\gamma_0} + \sum_{m\neq 0}\left(\frac{1}{2id\gamma_m} + \frac{1}{4\pi|m|}\right). \tag{4.86}$$

Using the third identity of eq.(4.81), for large values of m, the term in the sum is equal to:

$$\frac{1}{2id\gamma_m} + \frac{1}{4\pi|m|} \;\rightarrow\; -\frac{k^2}{4d|\alpha_m|^3} \;\rightarrow\; -\frac{d^2}{8\pi\lambda^2}\frac{1}{|m|^3}. \tag{4.87}$$

Obviously, the singularity and the slow convergence of $\mathcal{G}$ have been carried out by introducing the series $\mathcal{G}_\infty$. Fortunaley, this singularity is logarithmic and thus integrable, as shown in the next section.

Let us now deal with the second function defined in eq. (4.77):

$$\mathcal{N}(s,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}\left\{\frac{dx(s')}{ds'}\mathrm{sgn}\left[z(s)-z(s')\right] - \frac{\alpha_m}{\gamma_m}\frac{dz(s')}{ds'}\right\} e^{imK\left[x(s)-x(s')\right]+i\gamma_m|z(s)-z(s')|}. \tag{4.88}$$

At the first glance, the term $\mathrm{sgn}\left[z(s)-z(s')\right]$ suggests us that this function is not continuous for s' = s, at least if $\frac{dz'}{ds'} \neq 0$ for s' = s. To deal with this term, we proceed in the same way as in eq.(4.85), by introducing an asymptotic value $\mathcal{N}_\infty$ and we set now:

$$\mathcal{N} = \mathcal{N}_\infty + \left(\mathcal{N} - \mathcal{N}_\infty\right), \tag{4.89}$$

with

$$\begin{aligned}
\mathcal{N}_\infty(s,s') &= \frac{1}{2d}\,\mathrm{sgn}\left[z(s)-z(s')\right] \\
&+\frac{1}{2d}\left\{\frac{dx(s')}{ds'}\,\mathrm{sgn}\left[z(s)-z(s')\right]-i\frac{dz(s')}{ds'}\right\} e^{-\alpha_0\left[x(s)-x(s')\right]} \sum_{m=1}^{\infty} e^{-mK\left[x(s)-x(s')\right]+imK\left[z(s)-z(s')\right]} \\
&+\frac{1}{2d}\left\{\frac{dx(s')}{ds'}\,\mathrm{sgn}\left[z(s)-z(s')\right]+i\frac{dz(s')}{ds'}\right\} e^{\alpha_0\left[x(s)-x(s')\right]} \sum_{m=-1}^{-\infty} e^{mK\left[x(s)-x(s')\right]+imK\left[z(s)-z(s')\right]}.
\end{aligned} \tag{4.90}$$

The sums can be evaluated in a closed form:

$$\begin{aligned}
\mathcal{N}_\infty(s,s') &= \frac{1}{2d}\,\mathrm{sgn}\left[z(s)-z(s')\right] \\
&+\frac{1}{2d}\left\{\frac{dx(s')}{ds'}\,\mathrm{sgn}\left[z(s)-z(s')\right]-i\frac{dz(s')}{ds'}\right\} \frac{e^{-\alpha_0\left[x(s)-x(s')\right]}}{e^{K\left[x(s)-x(s')\right]-imK\left[z(s)-z(s')\right]}-1} \\
&+\frac{1}{2d}\left\{\frac{dx(s')}{ds'}\,\mathrm{sgn}\left[z(s)-z(s')\right]+i\frac{dz(s')}{ds'}\right\} \frac{e^{\alpha_0\left[x(s)-x(s')\right]}}{e^{K\left[x(s)-x(s')\right]+imK\left[z(s)-z(s')\right]}-1}.
\end{aligned} \tag{4.91}$$

As noticed for $\mathcal{G}$, the term $\mathcal{N}-\mathcal{N}_\infty$ must be considered as a series each term of which is obtained by making the difference of the corresponding terms in the sums in eqs.(4.88) and (4.90). This series converges much more rapidly than the series in eq. (4.88) and it is continuous at s = s'.

The limit of $\mathcal{N}$ when $s'\to s$ can be determined calculating the limits of the two terms $\mathcal{N}_\infty$ and $\left(\mathcal{N}-\mathcal{N}_\infty\right)$. After tedious calculations, we can deduce that this limit exists and is given by:

$$\mathcal{N}(s,s) = \lim_{s'\to s}\mathcal{N}(s,s') = -\frac{dz}{ds}\left(\frac{i}{2\pi\alpha_0}+\frac{1}{2d}\sum_{m=-\infty}^{\infty}\frac{\alpha_m}{\gamma_m}\right)+\frac{1}{4\pi}\frac{\dfrac{d^2z}{ds^2}}{\left(\dfrac{dx}{ds}\right)^2+\left(\dfrac{dz}{ds}\right)^2}. \tag{4.92}$$

This interesting results established by Pavageau and Bousquet[44] is very important for the numerical applications, because it shows that $\mathcal{N}(s,s)$ contains a series that canverges like $1/m^3$ (after adding the terms with negative and positive values of m).

Let us notice that the second derivative of the profile function appears in eq.(4.92) and it clearly requires the continuity of the first derivative, i.e., the absence of edges.

The third kernel $\mathcal{K}(s,s')$, given by eq. (4.36):

$$\mathcal{K}(s,s') = \frac{1}{2d}\sum_{m=-\infty,+\infty}\left[\mathrm{sgn}\left[z(s)-z(s')\right]\frac{dx}{ds}-\frac{\alpha_m}{\gamma_m^+}\frac{dz}{ds}\right]e^{imK\left[x(s)-x(s')\right]+i\gamma_m^+\left|z(s)-z(s')\right|} \tag{4.93}$$

is deduced from $\mathcal{N}(s,s')$ by replacing the derivatives $\frac{dx'}{ds'}$ and $\frac{dz'}{ds'}$ by $\frac{dx}{ds}$ and $\frac{dz}{ds}$ and its study is quite similar. After after tedious calculations it can be shown that:

$$\mathcal{K}(s,s)=\lim_{s'\to s}\mathcal{N}_0(s,s')=-\frac{dz}{ds}\left(\frac{i}{2\pi\alpha_0}+\frac{1}{2d}\sum_{m=-\infty}^{\infty}\frac{\alpha_m}{\gamma_m}\right)-\frac{1}{4\pi}\frac{\frac{d^2z}{ds^2}}{\left(\frac{dx}{ds}\right)^2+\left(\frac{dz}{ds}\right)^2}. \tag{4.94}$$

4.6.3. Integration of kernel singularities

Clearly, the asymptotic part of $\mathcal{G}_\infty(s,s')$ in eq.(4.84) is singular when $s\to s'$. After some calculations, it can be found that

$$\lim_{s\to s'}\mathcal{G}_\infty(s,s')=\frac{1}{4\pi}\ln\left(K^2\left\{\left[x(s)-x(s')\right]^2+\left[z(s)-z(s')\right]^2\right\}\right). \tag{4.95}$$

Noting that $x(s)-x(s')\approx(s-s')\frac{dx}{ds}$ and $z(s)-z(s')\approx(s-s')\frac{dz}{ds}$, eq.(4.95) yields:

$$\lim_{s\to s'}\mathcal{G}_\infty(s,s')=\frac{1}{2\pi}\left\{\ln(2\pi)+\frac{1}{2}\ln\left[\left(\frac{dx}{ds}\right)^2+\left(\frac{dz}{ds}\right)^2\right]+\ln\frac{|s-s'|}{d}\right\}. \tag{4.96}$$

The first two terms represent regular parts that can be integrated by using the rectangular rule, as shown later. Unfortunately, $\ln\frac{|s-s'|}{d}$ is not periodic and the rectangular rule is very poor when applied to nonperiodic functions. It is possible to overcome this difficulty [7] by considering another function defined on (0, d):

$$\tilde{\mathcal{G}}_\infty(s,s')=\frac{1}{2\pi}\left[\ln\frac{|s-s'|}{d}+\ln\left(1-\frac{|s-s'|}{d}\right)\right],\quad s,s'\in(0,d), \tag{4.97}$$

which has the same singularity as $\frac{1}{2\pi}\ln\frac{|s-s'|}{d}$ in the interval (0, d), because $d-|s-s'|$ never vanishes when $s,s'\in(0,d)$. The advantage of his new function is that it is continuous except on the singularity, and all its derivatives with respect to s are the same on $s' = 0$ and $s' = d$ and thus we can use the rectangular rule.

We perform the integration of $\mathcal{G}_\infty(s,s')$ by setting:

$$\int_0^{s_d}\mathcal{G}_\infty(s,s')\phi(s')ds'=\int_0^{s_d}\left[\mathcal{G}_\infty(s,s')\phi(s')-\tilde{\mathcal{G}}_\infty(s,s')\phi(s)\right]ds'+\int_0^{s_d}\tilde{\mathcal{G}}_\infty(s,s')\phi(s)ds'. \tag{4.98}$$

The term in the square bracket is non-singular and can be integrated using the rectangular rule, if we take into account that:

$$\lim_{s'\to s}\left[\mathcal{G}_\infty(s,s')\phi(s')-\tilde{\mathcal{G}}_\infty(s,s')\phi(s)\right]=\frac{1}{2\pi}\left\{\ln(2\pi)+\frac{1}{2}\ln\left[\left(\frac{dx}{ds}\right)^2+\left(\frac{dz}{ds}\right)^2\right]\right\}\phi(s). \tag{4.99}$$

The second term in eq.(4.98) contains the singular part, but it can be integrated analytically:

$$\int_0^{s_d} \tilde{\mathcal{G}}_\infty(s,s')\phi(s)ds' = \phi(s)\int_0^{s_d} \tilde{\mathcal{G}}_\infty(s,s')ds' = -\frac{d}{\pi}\phi(s)\,. \tag{4.100}$$

In conclusion, the integration of the singular kernel is made by introduction at first $\mathcal{G}_\infty$, which permits to define a series $\mathcal{G}-\mathcal{G}_\infty$ that is continuous and rapidly converging hence easily integrable by the use of the rectangular rule. Second, the integration of the term containing $\mathcal{G}_\infty$ is performed by defining a new function $\tilde{\mathcal{G}}_\infty$, which has the same singularity as $\mathcal{G}_\infty$, has the property of a periodic function, and can be analytically integrated.

4.6.4. Kernel singularities for highly conducting metals

When the conductivity of a metallic grating tends to infinity, the Green function in the metal tends to a delta function. This property is rather obvious: for very large conductivities, the field generated by a line current (delta function) placed in the metal decreases very rapidly since it is absorbed on very short distances. This behaviour have drastic consequences on the kernels of the integral equations dealing with metallic gratings, which are directly derived from the Green function: the two variable functions relative to the metallic part of the grating tend to delta functions as well. The integration of such functions through a point matching method requires more and more points of discretization around s′ = s and, since s can take any value in the interval (0, s_d) the integration and the inversion of the final linear system of equations (bearing in mind that its size is the total number of discretization points) becomes impossible. This remarks explains why the first attempts at implementing the integral equations on computers were not able to give any result for metallic gratings in the visible and infrared regions.

A very efficient way to overcome this difficulty is to apply an approach called *local summation* [7], using another form of the Green function [45]:

$$\mathcal{G}(\vec{r}-\vec{r}\,') = \frac{1}{4i}\sum_m e^{imd\alpha_0}\, H_0^+\left(k\left|\vec{r}-\vec{r}\,'\right| - md\,\hat{x}\right), \quad \vec{r} = (x,z), \tag{4.101}$$

with $\hat{x}$ being the unit vector of the x axis. This form is the direct consequence of the fact that $\frac{1}{4i}H_0^+\left(k\left|\vec{r}-\vec{r}\,'\right|\right)$ is the Green function of the Helmholtz equation:

$$\nabla^2\left[\frac{1}{4i}H_0^+\left(k\left|\vec{r}-\vec{r}\,'\right|\right)\right] + k^2\left[\frac{1}{4i}H_0^+\left(k\left|\vec{r}-\vec{r}\,'\right|\right)\right] = \delta\left(\vec{r}-\vec{r}\,'\right). \tag{4.102}$$

Since the pseudo-periodic Green function $\mathcal{G}(\vec{r}-\vec{r}\,')$ in vacuum is defined by:

$$\nabla^2\tilde{\mathcal{G}}(\vec{r},\vec{r}\,') + k^2\tilde{\mathcal{G}}(\vec{r},\vec{r}\,') = \sum_{m=-\infty,+\infty} e^{i\alpha_m d}\delta(\vec{r}-\vec{r}\,'-md\,\hat{x}), \tag{4.103}$$

it follows that $\mathcal{G}(\vec{r}-\vec{r}\,')$ is a sum of Green functions $\frac{1}{4i}H_0^+\left(k\left|\vec{r}-\vec{r}\,'\right|-md\,\hat{x}\right)$ satisfying:

$$\nabla^2\left[\frac{1}{4i}e^{i\alpha_n d}H_0^+\left(k\left|\vec{r}-\vec{r}\,'-md\,\hat{x}\right|\right)\right]+k^2\left[\frac{1}{4i}e^{i\alpha_n d}H_0^+\left(k\left|\vec{r}-\vec{r}\,'-md\,\hat{x}\right|\right)\right]= \\ =e^{i\alpha_n d}\delta\left(\vec{r}-\vec{r}\,'-md\,\hat{x}\right). \tag{4.104}$$

Inserting the value of $\mathcal{G}(\vec{r}-\vec{r}\,')$ given by eq. (4.101) inside the integral $\int_0^{s_d}\mathcal{G}(s,s')\phi(s')ds'$, then making the change of variable $\vec{r}\,'-md\,\hat{x}=\vec{r}\,''$, and finally gathering the infinite sum of integrals on one period into a single integral from $-\infty$ to $+\infty$ yields:

$$\int_0^{s_d}\mathcal{G}(s,s')\phi(s')ds'=\frac{1}{4i}\int_{-\infty}^{\infty}H_0^+(k\left|\vec{r}-\vec{r}\,'\right|)e^{i\alpha_0(x'-x)}\phi(s')ds'. \tag{4.105}$$

In the same way it can be shown that

$$\int_0^{s_d}\mathcal{N}(s,s')\psi(s')ds'=\frac{ik}{4}\int_{-\infty}^{\infty}H_1^+(k\left|\vec{r}-\vec{r}\,'\right|)e^{i\alpha_0(x'-x)}\frac{z'-z-\frac{dz}{ds'}(x'-x)}{\left|\vec{r}-\vec{r}\,'\right|}\psi(s')ds'. \tag{4.106}$$

When the permittivity becomes very large in modulus, the functions $\mathcal{G}^-(s,s')$ and $\mathcal{N}^-(s,s')$ are obtained by replacing k by kn^- in $\mathcal{G}(s,s')$ and $\mathcal{N}(s,s')$. Since the value of n^- is close to an imaginary number in the visible and infrared regions for usual metals (for example, n^- = 1.3 +i 7.11 for aluminum at 650 nm), the Hankel functions $H_0^+(k\left|\vec{r}-\vec{r}\,'\right|)$ and $H_1^+(k\left|\vec{r}-\vec{r}\,'\right|)$ become very close to the modified Bessel functions $K_0\left[k\left|n^-\left(\vec{r}-\vec{r}\,'\right)\right|\right]$ and $K_1\left[k\left|n^-\left(\vec{r}-\vec{r}\,'\right)\right|\right]$ (see [45]) and tend to delta functions when $\left|n^-\right|\to\infty$. Thus, these functions vary much more rapidly than the unknown function ϕ, which can be considered as a constant. Thus, remarking that when $x'\approx x$

$$\left|\vec{r}-\vec{r}\,'\right|\approx\left|x-x'\right|\sqrt{1+\left(\frac{dz}{dx}\right)^2}, \\ z'-z-(x'-x)\frac{dz}{dx}\approx-\frac{1}{2}\frac{d^2z}{dx^2}(x'-x)^2, \\ \exp\left[i\alpha_0(x'-x)\right]\approx 1, \tag{4.107}$$

yields finally:

$$\int_0^{s_d}\mathcal{G}(s,s')\phi(s')ds'\approx\frac{\phi(s)}{4i}\int_{-\infty}^{\infty}H_0^+(k\left|n^-\left(\vec{r}-\vec{r}\,'\right)\right|)e^{i\alpha_0(x'-x)}ds'\approx\frac{\phi(s)}{2ik\sqrt{1+\left(\frac{dz}{dx}\right)^2}}, \tag{4.108}$$

and

$$\int_0^{s_d} \mathcal{N}(s,s')\psi(s')ds' \approx \frac{\dfrac{d^2z}{dx^2}\psi(s)}{4ik\left[1+\left(\dfrac{dz}{dx}\right)^2\right]^{3/2}}. \qquad (4.109)$$

It is not surprising to note that in this approximation, the calculations of the integrals require neither summation of the kernels, nor integration of the singularities, nor matrix multiplication. As the kernels tend toward delta-distributions with amplitudes determining the coefficients in eqs.(4.108) and (4.109), their matrix representations tend to diagonal matrices.

Numerical results have shown that this simpler formulation not only successfully applies in the domain where the summation and integration processes defined in the previous sections fail, but also remains valid with a good accuracy (about 10^{-3} in relative value) in a large domain of metals and wavelengths. For example, with aluminum, this approach works in the visible, a domain in which the classical method of integration can be used as well (but with a greater computation time), wheras the local summation is necessary for metals in the far-infrared and microwaves domain. It is very important to notice that, using the local summation and assuming that $|n^-| \to \infty$, it can be shown that the integral equation for metallic gratings described in appendix 4.B tends towards the integral equations obtained for perfectly conducting metals.

4.6.5. Problems of edges and non-analytical profiles

When the grating profile has edges or corners (in 2D case), fundamental difficulties appear. First, the uniqueness of the solution of the electromagnetic field is not ensured. The hypothesis of the integrability of the unknown function u(x) in the integral equation is equivalent to the Meixner condition of integrability [46], although it is singular.

The second problem lies in the validity of the boundary condition of electromagnetic field on the grating profile. Indeed, on the edges, the normal and tangential direction on the profile are not defined in a unique manner. Moreover, when establishing these boundary conditions, the demonstration does not work if the surface has edges; However, these warnings are not dramatic. The demonstration of Archimede theorem is also questionable if the object presents edges. Does it exist any doubt about the validity of this theorem?

The third problem in the integral method lies in the process of integration in the vicinity of the edges. The kernels become discontinuous or even meaningless, depending on whether the point of calculation of the integral coincides with the edge point or not, see eq.(4.90) - (4.92). Moreover, the integration can fail due to the eventual singularity of the unknown function on the edge [44].

To overcome the edge problem, there exist several approaches. The most direct one consists in replacing the actual profile z = f(x) with its truncated Fourier representation:

$$\tilde{f}_M(x) = \sum_{m=-M}^{M} f_m e^{imKx}. \qquad (4.110)$$

The new profile has no edges and in most cases of ruled or holographic gratings, it mimics quite well the true profile. Numerical tests have shown a very good convergence of results with respect to the number 2M+1 of Fourier terms and the number P of discretization points, provided the empirical rule P > 4(2M + 1). This method can be used to describe some profiles that are not represented by continuous functions, for example lamellar gratings, provided that

M is larger than 10. Unfortanately, there are several important cases of classical gratings and classes of relatively new types of gratings and periodic systems that cannot be treated using this simple approach.

The first problem covers the case of echelle gratings, working in grazing incidence in high and very high orders (see Chapter 1). The fact that the period is some tens or hundreds times larger than the wavelength λ, and that the working facet is quite steep (sometimes going up to 86° groove angle), requires that its geometry is represented in the method with an error smaller than λ/20. Simple but *incorrect* estimations show that this rule of thumb would be acceptable for numerical treatment: if the period is close to 50 wavelengths, we need about 1000 points of discretization and thus 250 Fourier harmonics, according to the rule P > 4(2M + 1). However, Gibbs phenomenon will significantly modify the form of the working facet, which could be almost vertical. In order to correctly describe the field on this facet, we need to have at least 5 to 10 points per wavelength *along* the the working facet, rather than along its projection on the x-axis. With a 85° groove angle, the length of this facet is about 11 times its x-projection, in such a way that the number of discretization points must be multiplied by a factor 10. The number of Fourier components in the profile follows the same rule.

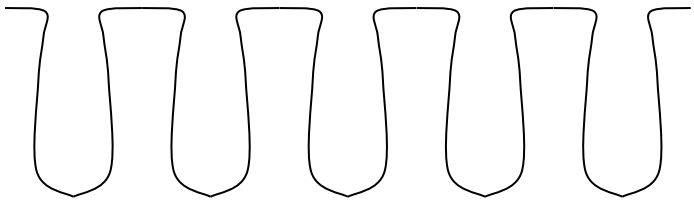


Figure 4.5. Schematic representation of an etched grating profile with non-analytical function description.

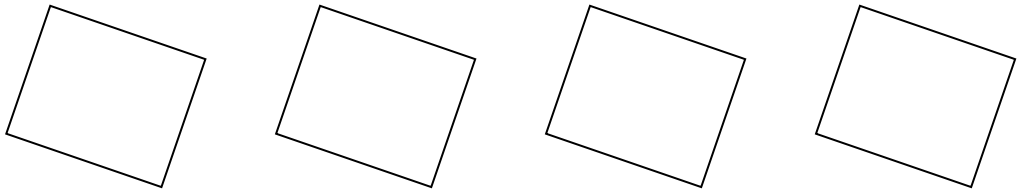


Figure 4.6. Grating made of inclined rectangular cross-section rods.

Another class of problems consists of unconventional geometries, like inverted slope grooves, obtained during groove etching technologies, as seen in Figure.4.5, or rod gratings, shown in Figure.4.6. We have noticed that the problem of vertical segments adds difficulties. For example, the problem of edges cannot be solved any longer by a Fourier expansion of the profile. It exists a possibility to simultaneously solve the two problems by introducing a curvilinear coordinate that follows the grating profile. As far as the integration is made along the profile, this curvilinear integration comes as a natural way of calculating the integrals in the integral equations. Moreover, adaptive meshing can be used to reduce the influence of edges.

In the real life, edges do not exist. On each edge there is at least one atom that has no edges, even though the light has a wavelength much larger than the atom dimensions. Edges of nuclear particles are not discussed even in the most exotic theories. Thus the idea is to replace the edges by arcs, where the partial derivatives can be well defined till the second order, which is sufficient for integral equations, as remarked in section 6.2. An adaptive

density of discretization points gives the possibility to significantly increase the density of the points close to the initial edges, and not elsewhere. Let us, for example, consider a rod grating having a straight rectangular cross section. The segments 1 and 3 are parallel to Ox, the segments 2 and 4, to Oz. Let us assume the origin of the curvilinear coordinate at the bottom-left corner (figure 4.7).

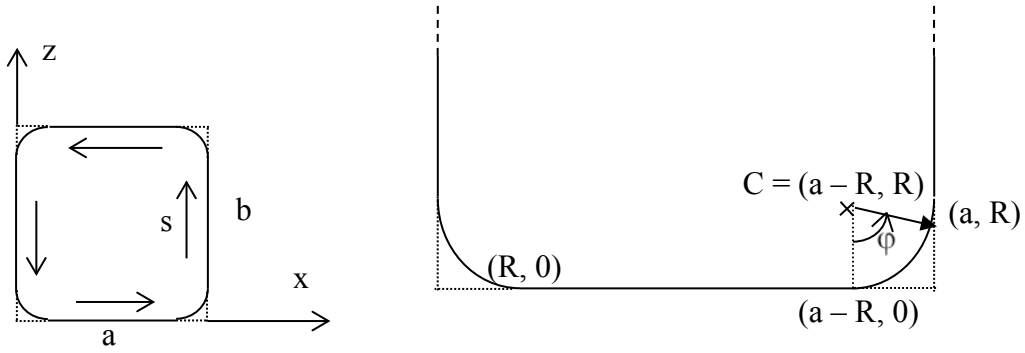


Figure 4.7. Left: rounding of corners of a rectangular cross-section rod with side lengths a and b, together with x, y, and s coordinate lines. Right: schematic representation of different straight and arc segments to obtain the links between the Cartesian and the curvilinear coordinates.

The coordinate s starts at x = R and z = 0, and follows the interface along its different parts:

1) the first horizontal segment, $0 < s < a - 2R$:

$$\begin{aligned} x(s) &= R + s, \\ z(s) &= 0. \end{aligned} \tag{4.111}$$

Here the final value of s is equal to $s_{1,max} = a - 2R$:

2) the circular rounding of the bottom-right corner, defined by the equation:

$$\left[x - (a - R)\right]^2 + \left(z - R\right)^2 = R^2. \tag{4.112}$$

Here, $0 < s - s_{1,max} < R\frac{\pi}{2}$:

$$\begin{aligned} s &= R\varphi + s_{1,max}, \\ x &= a - R + R\sin(\varphi) = a - R + R\sin\frac{s - s_{1,max}}{R}, \\ z &= R - R\cos(\varphi) = R - R\cos\frac{s - s_{1,max}}{R}. \end{aligned} \tag{4.113}$$

Here, $s_{2,max} = R\frac{\pi}{2} + s_{1,max}$

3) the next vertical segment at x = a, $0 < s - s_{2,max} < b - 2R$:

$$\begin{aligned} x &= a, \\ z &= s - s_{2,max} + R, \end{aligned} \tag{4.114}$$

and the same for the rest of the profile. The process is straightforward and needs an adapted application for each class of profiles. The advantage is that the derivatives $\frac{dx}{ds}$ and $\frac{dz}{ds}$ exist and are continuous everywhere on the profile. Thus, the second-order derivatives, which are required for the explicit summation and integration of the kernels, exist at least piesewisely, too. This could easily be checked at the point (a – R, 0), for example. For $s < s_{1,max}$, we use eqs.(4.111):

$$\frac{x(s)}{ds} = 1; \quad \frac{z(s)}{ds} = 0. \tag{4.115}$$

For $s_{1,max} < s < s_{2,max}$ it is necessary to use eqs.(4.113):

$$\begin{aligned} \frac{dx(s)}{ds} &= \cos\frac{s - s_{1,max}}{R} = 1 \quad \text{for } s = s_{1,max}, \\ \frac{dz(s)}{ds} &= \sin\frac{s - s_{1,max}}{R} = 0 \quad \text{for } s = s_{1,max}. \end{aligned} \tag{4.116}$$

The comparison of (4.115) and (4.116) points out the existence and continuity of the first derivatives.

It is worth noting that, in contrast with the adaptive *spatial* resolution used in several other methods (FEM, FDTD, RCW, coordinate transformation methods), here it is more convenient to call it adaptive *profile* resolution method, as far as it represents a 1D curvilinear coordinate adaptation.

Let us impose the requirement that the arc segments require N_{arc} times larger density points than on the straight segments. The entire length along s of the profile in figure.4.7 (left) is equal to $L_{tot} = 2a + 2b - 8R + 2\pi R$. The total number of discretization points is related to the length ot the segments, R, and to N_{arc}. If the distance between the points along the straight segments is Δ, it will be equal to Δ / N_{arc} on the arcs. Thus the total number of points for a single-rod per period is equal to:

$$P = \frac{2a + 2b - 8R + 2\pi R N_{arc}}{\Delta}. \tag{4.117}$$

In practice, unless some automatic scheme of determining the technical parameters of the computation is used, this equation determines Δ when the total number of integration points is chosen.

Starting from s = 0 in figure 4.8 (right), the abscissa values along s of the points on the interface are given by consecutively adding the values of the point number j to the end values of the previous segment:

1) first horizontal segment, $0 < s < s_{1,max}$: $s_j = s_{j-1} + \Delta$,
2) circular rounding of the bottom-right corner, $s_{1,max} < s < s_{2,max}$: $s_j = s_{j-1} + \frac{\Delta}{N_{arc}}$,
3) next vertical segment at x = a, $s_{2,max} < s < b - 2R + s_{2,max}$: $s_j = s_{j-1} + \Delta$,
4) so on to close the rod, then eventually going to another object inside the same grating period.

An additional improvement can be performed by making a smooth transfer from Δ to Δ/N_{arc}, i.e., to smoothly go from the density defined on the straight segments and on the arcs. This could be important for large-period systems with respect to the wavelength, if there

is a restriction in the total number of points in the profile discretization. If so, we can lay on the fact that the smaller the curvature of the segment (i.e., the larger its length), the smoother the behaviour of the kernels and of the unknown functions. Thus in the middle of a straight segment we will place points that are more distant to each other than close to the extremeties of the segments. Maystre has adapted such approach in his code Grating 2000, by defining a specific distance along the large segments starting from their edges, so that the density of the discretization points increases when approaching the ends of the segments. This approach gave the possibility to model echelle gratings working in very high diffraction orders (600 or more) in the '90s, when no alternative approach was able to provide reliable results. The method was unbeatable for echelles, before Li and Chandezon [47] formulated an improvement of the coordinate transformation method to work for profiles with edges. However, the latter does not apply to rod gratings, or to profiles having the form as given in figure.4.5.

4.7. Examples of numerical results

All the results shown in this section are obtained using the code Grating 2000.

4.7.1. Sinusoidal perfectly conducting grating

Table 1 shows the efficiencies in the two non-evanescent orders (-1st and 0th) of a sinusoidal perfectly conducting grating of period 600 nm and height 180 nm (from the bottom to the top of the groove) illuminated under incidence angle 30° and wavelength 600 nm. In this case, the -1st order is scattered with an angle of scattering (measured anticlockwise, in contrast with incident angle) of -30°, which entails that it propagates just in the direction which is the opposite to that of the incident wave (this is called Littrow mounting by specialists of gratings).

The number of discretisation points is P and the series included in the kernel are summed from –M to +M. The symbol $\sum \rho_m$ denotes the sum of the two efficiencies, the energy balance being satisfied when $\sum \rho_m = 1$.

	TE polarization			TM polarization		
P,M	ρ_{-1}	ρ_{-0}	$\sum \rho_m$	ρ_{-1}	ρ_{-0}	$\sum \rho_m$
4,2	0.4658	0.5437	1.0095	0.9466	0.0514	0.9980
6,3	0.4703	0.5288	0.9991	0.9581	0.0411	0.9992
25,10	0.4669	0.5336	1.0005	0.9579	0.0421	1.0000
50,20	0.4659	0.5334	0.9993	0.9579	0.0421	1.0000
110,50	0.4659	0.5341	1.0000	0.9579	0.0421	1.0000

Table 4.1. The sinusoidal perfectly conducting grating.

It is worth noting that a precision better than 0.01 is reached as soon as P > 4 and M > 2!. A precision of 10^{-3} needs P > 50 and M > 20.

4.7.2. Echelette perfectly conducting grating

The echelette grating is a grating with triangular groove (figure 4.8). The blaze angle b (angle of the large facet with the x axis) is equal to 30° and the apex angle A (angle between the two

facets) to 90°. The other parameters are the same as in section 7.1. It must be noticed that the incidence angle and the blaze angle are equal, which entails that the incident wavevector is orthogonal to the large facet. In these conditions, it can be shown that for TM polarization, the grating problem can be solved in closed form: the efficiency in the -1st order is equal to unity while in the 0th order it vanishes [48].

The demonstration is straightforward: the sum of the incident wave and of a plane wave with unit amplitude propagating in the opposite direction satisfies all the conditions of the boundary value problem stated in section 3.2. The reader can notice that this sum satisfies the Helmholtz equation. In addition, this sum of two plane waves propagating in opposite directions constitute an interference system that presents white areas and dark lines. The maximum of white lines coincide with the large facets of the grating, and on these lines, the derivative of the field (thus the normal derivative) vanishes. The normal derivative with respect to the small facet vanishes as well since the field is invariant in the normal direction. At the first glance, this property is obvious since the field is "reflected" by the large facets, or in other words, the scattering phenomenon reduces in that case to a simple reflection phenomenon. This reasoning fails since the same phenomenon is not observed for TE polarization: it is dangerous to invoke reflection phenomena on the large facets when the width of these facets has the same order of magnitude as the wavelength of the light! The concentration of the incident energy in a single order is called 'blazing effect' and the theorem of Marechal and Stroke is the origin of the name 'blazed gratings' given sometimes to ruled gratings.

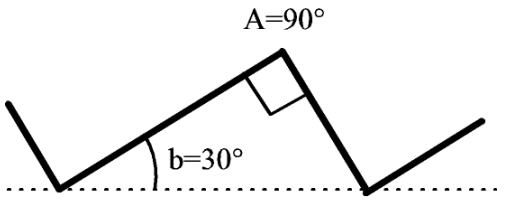


Figure 4.8: A ruled grating

	TE polarization			TM polarization		
P,M	ρ_{-1}	ρ_{-0}	$\sum\rho_m$	ρ_{-1}	ρ_{-0}	$\sum\rho_m$
6,3	0.6531	0.4451	1.0982	1.4452	0.0012	1.4464
25,10	0.5838	0.4123	0.9961	0.9976	0.0001	0.9977
50,20	0.5838	0.4123	0.9961	0.9976	0.0001	0.9977
110,50	0.5929	0.4055	0.9984	0.9984	0.0000	0.9984
250,100	0.5932	0.4035	0.9967	0.9989	0.0000	0.9989

Table 4.2. The echelette perfectly conducting grating and the Marechal and Stroke theorem.

Table 4.2 shows that the convergence is significantly slower than for the sinusoidal grating, due to the edges. Nevertheless, a precision of about 0.01 is obtained when P > 25 and M > 10. With the same values, the Marechal and Stroke theorem is satisfied with a precision better than 0.003. The computation time is always less than 1second on a PC computer except for the last line, for which 2 seconds are required.

4.7.3. Lamellar perfectly conducting grating

The profile of a lamellar grating is shown in figure 4.9. The widths of the hole and of the bump are denoted by t and b and the height by h. In this example, b = t = 300 nm, and h = 180 nm.

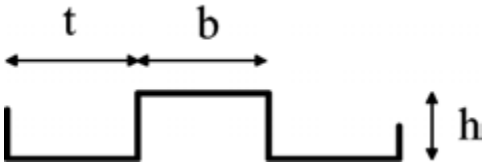


Figure 4.9. A lamellar grating

	TE polarization			TM polarization		
P,M	ρ_{-1}	ρ_{-0}	$\sum\rho_m$	ρ_{-1}	ρ_{-0}	$\sum\rho_m$
6,3	0.7770	0.6999	1.4769	23.66	5.09	28.75
25,10	0.3490	0.6394	0.9884	0.8293	0.1724	1.0016
50,20	0.3347	0.6569	0.9915	0.8191	0.1705	0.9896
150,70	0.3279	0.6659	0.9938	0.8201	0.1718	0.9919
250,100	0.3288	0.6733	1.0021	0.8214	0.1725	0.9939

Table 4.3. The echelette perfectly conducting grating and the Marechal and Stroke theorem

The main conclusion to draw from Table 4.3 is that the convergence for lamellar gratings is even slower than for echelette gratings. This is not surprising if we notice that the number of edges is multiplied by 2. A precision of about 0.02 is obtained when P > 50 and M > 20. The results for P = 6 and M = 3 are aberrant, specially for TM polarization.

4.7.4. Aluminum sinusoidal grating in the near infrared

We consider a sinusoidal aluminum grating with period d = 400 nm, a height h = 100 nm, illuminated with incidence angle 10° and wavelength 300 nm. With these parameters, three plane waves (-1^{st}, 0^{th} and $+1^{st}$) are reflected. The optical index of aluminum at 300 nm is equal to 4.2+ i 21.5. We give in Table 4.4 the efficiency in the -1^{st} order and the sum of the three efficiencies.

	TE polarization		TM polarization	
P,M	ρ_{-1}	$\sum\rho_m$	ρ_{-1}	$\sum\rho_m$
6,3	0.5205	0.9582	0.4367	0.9618
10,4	0.5201	0.9649	0.4325	0.9521
30,13	0.5203	0.9655	0.4321	0.9518
100,45	0.5204	0.9655	0.4320	0.9518

Table 4.4. The aluminum sinusoidal grating

Table 4.4 shows a very good convergence of the results, similar to the convergence observed for sinusoidal perfectly conducting gratings, thanks to the local summation of the two variable functions of the kernel derived from the Green function in aluminum.

4.7.5. Buried echelette silver grating in the visible.

The buried silver grating is shown in figure 4.10. A symmetric echelette silver grating with period 900 nm has been covered by a dielectric of index 1.5, the maximum depth e of dielectric being equal to 800 nm. This grating is illuminated in normal incidence .by a plane wave of wavelength 600 nm. The index of siver for this wavelength is equal to 0.006 + i 3.75. Table 4.5 gives the efficiency in the 0^{th} order and the total of efficiencies of the three reflected orders (-1^{st}, 0^{th} and $+1^{st}$).

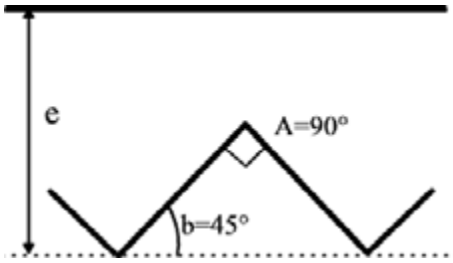


Figure 4.10. A symmetric silver grating coated by dielectric

	TE polarization		TM polarization	
P,M	ρ_0	$\sum\rho_m$	ρ_0	$\sum\rho_m$
10,4	0.4892D	0.9830	0.6316	1.0622
30,13	0.5164	0.9651	0.5665	0.9339
100,45	0.5153	0.9600	0.6062	0.9178
200,90	0.5153	0.9593	0.6168	0.9153

Table 4.5. The buried silver grating

The convergence is slower than in the preceding case and the results for P = 10 and M = 4 are not correct, specially for TM polarization. For TE polarization, a convergence of the results with a precision better than 0.006 is obtained for P = 30 and M = 13. This is not so for TM polarization, in which it is necessary to reach P = 100 and M = 45 to get a precision of the order of 0.003. Here, the method of local summation is not used, but this fact does not explain the slower convergence since the modulus of the optical index is not large. The main reason can be found in the edges of the profile. The slower convergence for TM polarization is rather general for metallic gratings, due to the existence of plasmon resonances on the grating surface.

4.7.6. Dielectric rod grating.

The grating is made of dielectric elliptic rods with optical index 1.4, width w = 600 nm and height h = 400 nm (figure 4.11). The period is equal to 800 nm. It is illuminated with incidence 20° by a plane wave with wavelength 600 nm. Two orders (-1^{st} and 0^{th}) are reflected and transmitted. We give in Table 4.6 the efficiency in the 0^{th} transmitted order and

the sum of efficiencies of the 4 scattered orders, which should be equal to 1 for a perfect energy balance.

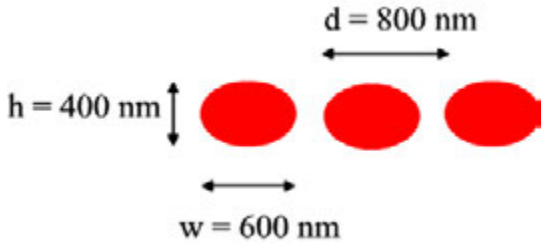


Figure 4.11. A dielectric rod grating

	TE polarization		TM polarization	
P,M	τ_0	$\sum\rho_m + \sum\tau_m$	τ_0	$\sum\rho_m + \sum\tau_m$
10,4	0.1824D	0.9139	0.7893	1.2489
30,13	0.2163	0.9961	0.8161	1.0054
100,45	0.2073	1.0005	0.8187	1.0004
200,90	0.2070	1.0001	0.8187	1.0001

Table4. 6. The dielectric rod grating with elliptic rods

The precision for P = 30 and M = 13 is equal to 0.01 and for P = 100 and M = 90, it reaches 0.0004.

4.7.7. Flat perfectly conducting rod grating

The grating is similar to that of figure 4.11, but its height is very small (8 nm). In order to obtain a significant reflection, the dielectric has been replaced by a perfectly conducting material

	TE polarization		TM polarization	
P,M	τ_0	$\sum\rho_m + \sum\tau_m$	τ_0	$\sum\rho_m + \sum\tau_m$
100,45	0.0598	1.1414	0.8940	0.9759
200,90	0.0634	1.0086	0.0292	0.9887
300,140	0.0639	1.0006	0.1059	0.9996
400,190	0.0639	1.0000	0.1119	1.0000

Table4.7. The flat perfectly conducting rod grating.

The convergence is very slow and it is necessary to reach P = 300 and M = 140 to obtain a precision better than 0.006, the energy balance being satisfied with a precision better than 10^{-3}. The reason must be found in the small height of the rods, as explained in section 6.1. The functions included in the kernel of the integral equation become very large in

modulus and very small in width for two points located on both sides of the rod for the same abscissa, thus the integration needs a large density of discretization points.

Appendix 4.A. Mathematical bases of the integral theory

4.A.1. Presentation of the mathematical problem

We consider (figure 4.1) a function $F(x,z)=\begin{cases} F^+(x,z) \text{ in } V^+, \\ F^-(x,z) \text{ in } V^-, \end{cases}$ which satisfies the following conditions:

- it is pseudo-periodic along the x axis:

$$F(x+d,z) = F(x,z)\exp(i\alpha_0 d), \tag{4.118}$$

- it satisfies a Helmholtz equation:

$$\nabla^2 F^{\pm} + k^2\left(n^{\pm}\right)^2 F^{\pm} = 0 \quad \text{in } V^{\pm}, \tag{4.119}$$

- it satisfies a radiation condition for $z \to \pm\infty$.

The aim of this appendix is to use the second Green's identity and basic theorems on boundary value problems in order to find an integral expression of this function and to analyze the properties of this integral expression. We will deduce the basic keys for writing an integral equation from a boundary value problem.

4.A.2. Calculation of the Green function

The first step of the calculation is to find the pseudo-periodic elementary solutions $\mathcal{G}^+(\vec{r})$ and $\mathcal{G}^-(\vec{r})$ of the two Helmholtz equations condensed in eq. (4.119) (with constants $k^2(n^+)^2$ for $\mathcal{G}^+$ and with constant $k^2(n^+)^2$ for $\mathcal{G}^-$, which satisfy the radiation conditions for $z \to \pm\infty$ and the following equations :

$$\nabla^2\mathcal{G}^+(\vec{r}) + k^2\left(n^+\right)^2\mathcal{G}^+(\vec{r}) = \sum_{m=-\infty,+\infty} e^{i\alpha_m d}\delta(\vec{r} - md\,\hat{x}), \text{ in } V^+ \text{ and } V^-, \tag{4.120}$$

$$\nabla^2\mathcal{G}^-(\vec{r}) + k^2\left(n^-\right)^2\mathcal{G}^-(\vec{r}) = \sum_{m=-\infty,+\infty} e^{i\alpha_m d}\delta(\vec{r} - md\,\hat{x}), \text{ in } V^+ \text{ and } V^-, \tag{4.121}$$

$$\mathcal{G}^{\pm}(\vec{r} + d\,\hat{x}) = \mathcal{G}^{\pm}(\vec{r})e^{i\alpha_0 d}, \tag{4.122}$$

with:

$$\alpha_m = \alpha_0 + mK, \qquad K = \frac{2\pi}{d}, \tag{4.123}$$

and $\hat{x}$ being the unit vector of the x axis.

We must emphasize that, in contrast with eq. (4.119), in which F^+ and F^- do not satisfy the same Helmholtz equation, each Green function satisfies a unique Helmholtz equation in the entire space, with constant $k^2(n^+)^2$ for $\mathcal{G}^+$ and with constant $k^2(n^+)^2$ for $\mathcal{G}^-$.

After expanding the periodic function $\mathcal{G}^{\pm}\, e^{-i\alpha_0 x}$ in Fourier series, then multiplying the Fourier series by $e^{i\alpha_0 x}$, it can be deduced that:

$$\mathcal{G}^{\pm}(x,z) = \sum_{m=-\infty,+\infty} G^{\pm}{}_m(z)\, e^{i\alpha_m x} . \tag{4.124}$$

On the other hand, the right-hand member of eq. (4.120) , called Dirac comb, can also be expanded in series:

$$\sum_{m=-\infty,+\infty} e^{i\alpha_m d}\delta(\vec{r} - nd\hat{x}) = \frac{1}{d}\delta(z)\sum_m e^{i\alpha_m x} . \tag{4.125}$$

Introducing equations (4.124) and (4.125) in equation (4.120), multiplying by $e^{-i\alpha_0 x}$ then identifying the coefficients of the Fourier series yield:

$$G_m^{\pm}{}''(z) + \left(\gamma_m^{\pm}\right)^2 G_m^{\pm}(z) = \frac{1}{d}\delta(z) , \tag{4.126}$$

with $\left(\gamma_m^{\pm}\right)^2 = k^2\left(n^{\pm}\right)^2 - \alpha_m^2$.

For $z \neq 0$, eq.(4.126) becomes the well-known one-dimension propagation equation in a homogeneous media without sources and have for solutions exponentials. We are searching for plane waves that satisfy the radiation condition for $z \to \pm\infty$, thus:

$$G_m^{\pm}(z) = \begin{cases} A_m^{\pm} \exp\left[i\gamma_m^{\pm} z\right], & \text{if } z > 0, \\ B_m^{\pm} \exp\left[-i\gamma_m^{\pm} z\right], & \text{if } z < 0. \end{cases} \tag{4.127}$$

Distribution theory proves that the solution $G_m^{\pm}$ of eq.(4.126) is a continuous function of z and that its derivative has a jump equal to 1/d at $z = 0$. These two conditions allow one to find the unknown amplitudes:

$$\begin{aligned} A_m^{\pm} &= B_m^{\pm}, \\ i\gamma_m^{\pm} A_m^{\pm} - \left(-i\gamma_m^{\pm} B_m^{\pm}\right) &= \frac{1}{d}, \end{aligned} \tag{4.128}$$

and thus:

$$G_m^{\pm}(z) = \frac{1}{2i\gamma_m^{\pm} d}\exp\left(i\gamma_m^{\pm}|z|\right). \tag{4.129}$$

The final form of $\mathcal{G}^{\pm}$ becomes:

$$\mathcal{G}^{\pm}(\vec{r}) = \frac{1}{2id}\sum_m \frac{1}{\gamma_m^{\pm}}\exp\left[i\alpha_m x + i\gamma_m^{\pm}|z|\right]. \tag{4.130}$$

Once the source is not in the origin, the Green functions is the function $\mathcal{G}^{\pm}(\vec{r} - \vec{r}\,')$. Notice that $\mathcal{G}^{\pm}$ symbolizes the Green functions.

4.A.3. Integral expression

Now, we apply the second Green theorem in order to find the expression of the function $F^{\pm}$. First, we consider the expression of F^{-} in V^{-}:

$$F^{-}(x,z) = \int_{S_T} \mathcal{G}^{-}(x-x',z-z')\frac{dF^{-}(x',z')}{dN_S}ds' - \int_{S_T} \frac{d\mathcal{G}^{-}(x-x',z-z')}{dN_S}F^{-}(x',z')ds', \qquad (4.131)$$

with the normal $\vec{N}_S$ being oriented towards the exterior of V^{-}, $x = x(s)$, $x' = x(s')$, and similar expressions for z and z′. The curve S_T includes four parts: the vertical lines S_L at x = 0 and S_R at x = d, the horizontal segment S_H at $z = z_H < 0$ (figure 4.12), and, finally, one period of S. The variable s′ denotes the curvilinear abscissa on S_T^{-}, with origine being located at the origin of the Cartesian coordinates.

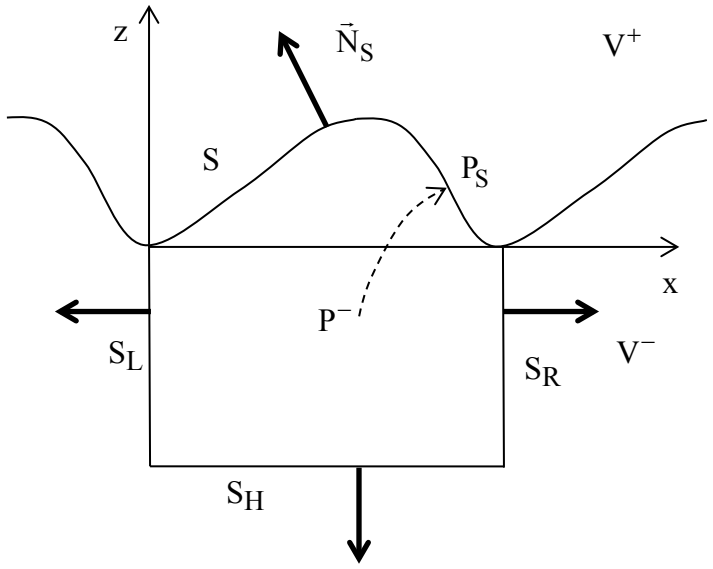


Figure 4.12. Application of the second Green theorem.

The pseudo-periodicity in x′ of $F^{-}(x',z')$ (and of $dF^{-}(x',z')/dN_S$) is opposite to that of $\mathcal{G}^{-}(x-x',z-z')$ (or of $d\mathcal{G}^{-}(x-x',z-z')/dN_S$), which entails that $\mathcal{G}^{-}(x-x',z-z')\frac{dF^{-}(x',z')}{dN_S}$ and $\frac{d\mathcal{G}^{-}(x-x',z-z')}{dN_S}F^{-}(x',z')$ are periodic. Furthermore, taking into account the orientation of the normal on S_R and S_L, $\frac{dF^{-}(x',z')}{dN_s} = -\frac{dF^{-}(x',z')}{dx}$ and $\frac{d\mathcal{G}^{-}(x-x',z-z')}{dN_S} = -\frac{d\mathcal{G}^{-}(x-x',z-z')}{dx}$ on S_L, while $\frac{dF^{-}(x',z')}{dN_s} = +\frac{dF^{-}(x',z')}{dx}$ on S_R. Thus, the integrals on S_R and S_L in eq. (4.131) cancel each other. On S_H, ds′ and dx′ identify and the integral takes the form:

$$\int_{S_H}\left[\mathcal{G}^-\left(x-x',z-z'\right)\frac{dF^-(x',z')}{dN_s}-\frac{d\mathcal{G}^-\left(x-x',z-z'\right)}{dN_S}F^-(x',z')\right]ds'$$

$$=\frac{1}{2d}\sum_{m=-\infty}^{\infty}e^{i\gamma_m^-(z-z_H)}\int_0^d e^{imK(x-x')}\left[\frac{1}{i\gamma_m^-}\phi_H^-(s')+\psi_H^-(s')\right]dx' \qquad (4.132)$$

$$=\frac{1}{2}\sum_{m=-\infty}^{\infty}e^{i\gamma_m^-(z-z_H)+imKx}\left(\frac{1}{i\gamma_m^-}\phi_{H,m}^-+\psi_{H,m}^-\right),$$

where $\phi_{H,m}^-$ and $\psi_{H,m}^-$ are the m^{th} Fourier components of the periodic functions $\phi_H^-\left(s'\right)=\frac{dF^-(x',z')}{dN_s}e^{-i\alpha_0x'}$ and $\psi_H^-\left(s'\right)=F^-(x',z')e^{-i\alpha_0x'}$ defined on S_H. Here, we can take advantage of the fact that the segment is parallel to the x axis and lies outside the groove region, which entails that $\phi_{H,m}^-$ and $\psi_{H,m}^-$ are related through the plane wave expansion valid in the homogeneous region (see chapter 2):

$$F^-=\sum_{m=-\infty,+\infty}t_me^{i\alpha_mx-i\gamma_m^-z}\quad \text{if } z<0, \qquad (4.133)$$

which yields:

$$\psi_{H,m}^-=-i\gamma_m^-\phi_{H,m}^-. \qquad (4.134)$$

This relation allows us to cancel the integral along S_H and thus the values of $F^-(x,z)$ can be determined by an integral along a single groove of S:

$$F^-\left(x,z\right)=\int_{s'=0}^{s_d}\mathcal{G}^-\left(x-x',z-z'\right)\frac{dF^-(x',z')}{dN_s}ds'-\int_{s'=0}^{s_d}\frac{d\mathcal{G}^-\left(x-x',z-z'\right)}{dN_S}F^-(x',z')ds', \qquad (4.135)$$

with s_d being the curvilinear abscissa of the point of S of abscissa d, the origin of curvilinear abscissa being the origin of the Cartesian coordinates system.

Similar considerations apply to $F^+(x,z)$, so that, after elementary calculations:

$$F^\pm(x,z)=\pm\int_{s'=0}^{s_d}\left[\mathcal{G}^\pm(x,z,s')e^{i\alpha_0(x-x')}\phi^\pm(s')e^{i\alpha_0x'}+\mathcal{N}^\pm(x,z,s')e^{i\alpha_0(x-x')}\psi^\pm(s')e^{i\alpha_0x'}\right]ds' \qquad (4.136)$$

with:

$$\mathcal{G}^\pm(x,z,s')=\mathcal{G}^\pm(x,z,x\left(s'\right),z\left(s'\right))=\frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_m^\pm}e^{imK\left(x-x'(s')\right)+i\gamma_m^\pm\left|z-z'(s')\right|}, \qquad (4.137)$$

$$\mathcal{N}^{\pm}(x,z,s') = -\frac{\partial \mathcal{G}^{\pm}}{\partial N_s}(x,z,x'(s'),z'(s')) = \\ = \frac{1}{2d}\sum_{m=-\infty}^{\infty}\left[\frac{dx'}{ds'}\operatorname{sgn}(z-z') - \frac{\alpha_m}{\gamma_m^{\pm}}\frac{dz'}{ds'}\right] e^{imK(x-x')+i\gamma_m^{\pm}|z-z'|}, \tag{4.138}$$

with $\psi^{\pm}(s') = F^{-}(x',z')e^{-i\alpha_0 x'}$ and $\phi^{\pm}(s') = \frac{dF^{-}(x',z')}{dN_s}e^{-i\alpha_0 x'}$ being defined on the grating profile S. Defining the periodic function $U^{\pm}(x,z) = F^{\pm}(x,z)e^{-i\alpha_0 x}$ yields:

$$U^{\pm}(x,z) = \pm\int_{s'=0}^{s_d}\left[\mathcal{G}^{\pm}(x,z,s')\phi^{\pm}(s') + \mathcal{N}^{\pm}(x,z,s')\psi^{\pm}(s')\right]ds'. \tag{4.139}$$

4.A.4. Equation of compatibility

In this section, we establish a crucial property of the integral theory of gratings, which unfortunately is ignored in most of the reference books of Electromagnetics. With this aim, it is necessary to point out a fundamental property of the expression of $U^{\pm}(x,z)$ given by eqs. (4.13) and (4.139). Let us suppose that we introduce in eq. (4.136) arbitrary periodic functions $\tilde{\phi}^{-}(s')$ and $\tilde{\psi}^{-}(s')$. Since we have not introduced the actual physical values $\phi^{-}(s')$ and $\psi^{-}(s')$ of these functions, we cannot expect to obtain the actual value of $F^{-}(x,z)$, but another function $\tilde{F}^{-}(x,z)$. More important, if the point $P^{-}(x,z)$ of V^{-} tends towards a point P_S of curvilinear abscissa s located on the profile (figure 4.12), the limit of $\tilde{F}^{-}(x,z)$ below the profile, denoted $\lim_{-}\left\{\tilde{F}^{-}(x,z)\right\}$ is not equal to $\tilde{\psi}^{-}(s')e^{i\alpha_0 x'}$. The same remark can be made for the normal derivative $\lim_{-}\left\{\frac{d\tilde{F}^{-}(x,z)}{dN_s}\right\}$, which is different from $\tilde{\phi}^{-}(s')e^{i\alpha_0 x'}$. In order to understand this surprising property, it is necessary to give two results which can be demonstrated using the theory of distributions. Let us give these two fundamental results without demonstration.

1. The integral expression of $\tilde{F}^{-}(x,z)$ inside V^{-} satisfies the Helmholtz equation in V^{-} like the actual field. What happens to the same integral expression of $\tilde{F}^{-}(x,z)$, but now calculated in region V^{+}? Denoting by $\tilde{F}^{-}_{ext}(x,z)$ the function equal to $\tilde{F}^{-}(x,z)$ in V^{-} and to this integral extension in V^{+}, it is easy to verify that $\tilde{F}^{-}_{ext}(x,z)$ satisfies in the entire space the Helmholtz equation satisfied by $\tilde{F}^{-}(x,z)$ in V^{-}, with constant $k^2(n^{-})^2$.

2. The jumps $\lim_{+}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}-\lim_{-}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}$ and $\lim_{+}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}-\lim_{-}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}$ of $\tilde{F}_{ext}^{-}(x,z)$ and of its normal derivative across S (difference between the values in V^{+} and in V^{-}) are respectively equal to $-\tilde{\psi}^{-}(s')e^{i\alpha_0 x'}$ and $-\tilde{\phi}^{-}(s')e^{i\alpha_0 x'}$.

The conclusion to draw from these properties is that the limit of $\tilde{F}^{-}(x,z)$ and of its normal derivative on S are equal to $\tilde{\psi}^{-}(s')e^{i\alpha_0 x'}$ and $\tilde{\phi}^{-}(s')e^{i\alpha_0 x'}$ in one case only: if $\tilde{F}_{ext}^{-}(x,z)$ vanishes throughout V^{+}. Indeed, if this property is satisfied, the jumps are nothing but $-\lim_{-}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}$ and $-\lim_{-}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}$, thus $\lim_{-}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}=\tilde{\psi}^{-}(s)e^{i\alpha_0 x}$ and $\lim_{-}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}=\tilde{\phi}^{-}(s)e^{i\alpha_0 x}$. This case occurs if the values of $\tilde{\psi}^{-}(s')$ and $\tilde{\phi}^{-}(s')$ introduced in the integral expression are equal to the actual physical values $\psi^{-}(s')$ and $\phi^{-}(s')$.

This result is not surprising for the specialist of boundary value problems. Indeed, using the second Green theorem in V^{-}, we introduce in the integral expression of the field the limit values of both the field and of its normal derivative below S. In the domain of boundary value problems, it is well known that *if a function must satisfy a Helmholtz equation in* V^{-} *and a radiation condition for* $z\to-\infty$*, one cannot impose the limit values of both this function and its normal derivative on S. In fact, we can impose either the limit values of this function or that of its normal derivative on S: in both cases, the solution of the boundary value problem exists and is unique***.** Unfortunately, it does not exist any tool of applied mathematics which enables one to express the field in an integral form including either the limit values of the field, or that of its normal derivative on S: a prior solution of an integral equation is required, which is much more difficult.

On the other hand, if both the actual values of the field and of its normal derivative are known, the field can directly be expressed in an integral form through the second Green theorem, without any integral equation, and this is why the second Green theorem is considered as a basic tool of the integral mathods of scattering. It must be emphasized that in that case, we know *a priori* that the the limit values of the field and of its normal derivative are the actual ones, thus that they are compatible, which is not the case for arbitrarily chosen limits.

This fundamental property shows that when the second Green theorem is used with unknown values of the limits of the field and of its normal derivative, we must impose to these limits an equation of compatibility. This compatibility is satisfied if we impose to the limit $\lim_{-}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}$ of the integral expression of $\tilde{F}_{ext}^{-}(x,z)$ to be equal to $\tilde{\psi}^{-}(s)e^{i\alpha_0 x}$ or if we impose to the limit $\lim_{+}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}$ of the integral expression of $\tilde{F}_{ext}^{-}(x,z)$ to be equal to zero, these two conditions being equivalent. Indeed, if $\lim_{+}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}=0$, the expression of $\tilde{F}_{ext}^{-}(x,z)$ in V^{+} satisfies a Helmholtz equation, a radiation condition at infinity and its limit

on S vanishes. The obvious solution of this boundary value problem is $\tilde{F}_{ext}^{-}(x,z)=0$ in V^{+}, and we have seen that the solution of this boundary value problem is unique, thus it is the solution. The consequence is that $\lim_{+}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}=0$, thus $\lim_{-}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}=\tilde{\phi}^{-}(s)e^{i\alpha_0 x}$.

In order to implement this condition of compatibility, we have to express $\lim_{+}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}$ or $\lim_{-}\left\{\frac{d\tilde{F}_{ext}^{-}(x,z)}{dN_s}\right\}$. We can use a first equation:

$$\lim_{+}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}-\lim_{-}\left\{\tilde{F}_{ext}^{-}(x,z)\right\}=-\tilde{\psi}^{-}(s)e^{i\alpha_0 x}, \tag{4.140}$$

or equivalently:

$$\lim_{+}\left\{\tilde{U}_{ext}^{-}(x,z)\right\}-\lim_{-}\left\{\tilde{U}_{ext}^{-}(x,z)\right\}=-\tilde{\psi}^{-}(s). \tag{4.141}$$

In order to find a second equation, we can consider eq. (4.139) which gives the expression of $U^{-}(x,z)=F^{-}(x,z)e^{-i\alpha_0 x}$. If z is fixed, this expression is a Fourier series in x, which is discontinuous on S. It is well known that the value on S of this Fourier series is the average value of the limits on both sides of S, thus:

$$\lim_{+}\left\{\tilde{U}_{ext}^{-}(x,z)\right\}+\lim_{-}\left\{\tilde{U}_{ext}^{-}(x,z)\right\}=-2\int_{s'=0}^{s_d}\left[\mathcal{G}^{-}(s,s')\tilde{\phi}^{-}(s')+\mathcal{N}^{-}(s,s')\tilde{\psi}^{-}(s')\right]ds', \tag{4.142}$$

with $\mathcal{G}^{-}(s,s')$ and $\mathcal{N}^{-}(s,s')$ being the integral expressions of $\mathcal{G}^{-}(x,z,s')$ and $\mathcal{N}^{-}(x,z,s')$ when the point of coordinates x,z becomes a point of S of curvilinear abscissa s:

$$\mathcal{G}^{-}(s,s')=\mathcal{G}^{-}(x(s),z(s),s'), \tag{4.143}$$

$$\mathcal{N}^{-}(s,s')=\mathcal{N}^{-}(x(s),z(s),s'). \tag{4.144}$$

From eqs. (4.141) and (4.142), we deduce the limits of $\tilde{U}^{-}$ on both sides of S, and we derive the equation of compatibility in V^{-}:

$$\int_{s'=0}^{s_d}\left[\mathcal{G}^{-}(s,s')\tilde{\phi}^{-}(s')+\mathcal{N}^{-}(s,s')\tilde{\psi}^{-}(s')\right]ds'+\frac{\tilde{\psi}^{-}(s)}{2}=0. \tag{4.145}$$

Achieving a similar calculation for V^{+} yields the general equation of compatibility in $V^{\pm}$:

$$\int_{s'=0}^{s_d}\left[\mathcal{G}^{\pm}(s,s')\tilde{\phi}^{\pm}(s')+\mathcal{N}^{\pm}(s,s')\tilde{\psi}^{\pm}(s')\right]ds'\mp\frac{\tilde{\psi}^{\pm}(s)}{2}=0, \tag{4.146}$$

$$\mathcal{G}^{\pm}(s,s')=\frac{1}{2id}\sum_{m=-\infty}^{\infty}\frac{1}{\gamma_m^{\pm}}e^{imK(x(s)-x'(s'))+i\gamma_m^{\pm}|z(s)-z'(s')|}, \tag{4.147}$$

$$\mathcal{N}^{\pm}(s,s') = \frac{1}{2d}\sum_{m=-\infty}^{\infty}\left[\frac{dx'}{ds'}\operatorname{sgn}\left(z(s)-z'(s')\right)-\frac{\alpha_m}{\gamma_m^{\pm}}\frac{dz'}{ds'}\right]e^{imK\left(x(s)-x'(s')\right)+i\gamma_m^{\pm}\left|z(s)-z'(s')\right|}. \quad (4.148)$$

It can be shown [7] that $\mathcal{G}^{\pm}(s,s')$ has an integrable logarithmic singularity (it behaves like $a_0 + a_1 \mathrm{Log}|s-s'|$ when $s' \to s$, a_0 and a_1 complex numbers) which can be taken into account in the integral by removing the singularity $a_1 \mathrm{Log}|s-s'|$ from $\mathcal{G}^{\pm}(s,s')$ then by integrating it in closed form. At the first glance, $\mathcal{N}^{\pm}(s,s')$ cannot be continuous, due to the discontinuity of $\operatorname{sgn}\left(z(s)-z'(s')\right)$ for s=s'. In fact, a careful analysis of this function around s = s' shows that it is continuous and that its limit when $s' \to s$ can be expressed in closed form [44,7], as stated in section 6.3.

4.A.5. Generalized compatibility

In the physical problem, the total field in V^+ includes the incident field. Here, we give an extension of the compatibility equation to the total field which allows a significant simplification of the use of integral theory.

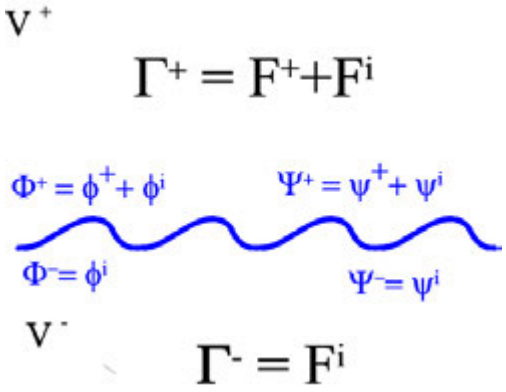


Figure 4.13. generalized: generalized compatibility

We define (figure 4.13) a function $\Gamma = \begin{cases} \Gamma^+ = F^{T+} = F^+ + F^i \text{ in } V^+, \\ \Gamma^- = F^i \text{ in } V^-. \end{cases}$

F^+ being the field scattered in V^+ by a grating illuminated by the incident field F^i. Γ^- is the expression of the incident field in V^-. Thus, in contrast with the preceding section, the function considered in this section does not satisfy a radiation condition at infinity in V^+. According to eq. (4.136), the value of F^+ is given by:

$$\begin{aligned} F^+(x,z) = \int_{s'=0}^{s_d} \Big[& \mathcal{G}^+(x,z,s')e^{i\alpha_0(x-x')}\phi^+(s')e^{i\alpha_0 x'} \\ & + \mathcal{N}^+(x,z,s')e^{i\alpha_0(x-x')}\psi^+(s')e^{i\alpha_0 x'}\Big]ds', \end{aligned} \quad (4.149)$$

thus Γ^+ can be written:

$$\Gamma^+(x,z) = F^i(x,y) + \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \phi^+(s') e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \psi^+(s') e^{i\alpha_0 x'} \right] ds'. \tag{4.150}$$

We denote by $\Psi^+ = \psi^+ + \psi^i$ and $\Phi^+ = \phi^+ + \phi^i$ the limits of Γ^+ on S and its normal derivative respectively. Introducing these values in eq. (4.149) yields:

$$F^+(x,z) = \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \left(\Phi^+(s') - \phi^i(s') \right) e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \left(\Psi^+(s') - \psi^i(s') \right) e^{i\alpha_0 x'} \right] ds'. \tag{4.151}$$

Thus, the integral expression of the total field Γ^+ is given by:

$$\Gamma^+(x,z) = F^i(x,z) + \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \left(\Phi^+(s') - \phi^i(s') \right) e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \left(\Psi^+(s') - \psi^i(s') \right) e^{i\alpha_0 x'} \right] ds', \tag{4.152}$$

or, gathering the terms containing the incident field:

$$\Gamma^+(x,z) = F^i(x,z) - \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \phi^i(s') e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \psi^i(s') e^{i\alpha_0 x'} \right] ds' \tag{4.153}$$

$$+ \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \Phi^+(s') e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \Psi^+(s') e^{i\alpha_0 x'} \right] ds'.$$

In order to simplify this equation, we consider the part of the integral containing the incident field, the middle line in the equation above. This part is equal to 0. Indeed, we know that $\phi^i(s')$ and $\psi^i(s')$ are compatible in V^- since they are derived from the actual values of the incident field in V^- (figure 4.13). We have shown in the preceding section that the integral expression of such a function vanishes in V^+. Remarking that $\mathcal{G}^-(x,z,s') = \mathcal{G}^+(x,z,s')$ and $\mathcal{N}^-(x,z,s') = \mathcal{N}^+(x,z,s')$ since Γ^+ and Γ^- satisfy the same Helmholtz equation (with constant $k^2(n^+)^2$), we can write than in V^+:

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0(x-x')} \phi^i(s') e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0(x-x')} \psi^i(s') e^{i\alpha_0 x'} \right] ds' = 0. \tag{4.154}$$

Thus finally the expression of Γ^+ is given by:

$$\Gamma^{+}(x,z) = F^{i}(x,z) + \int_{s'=0}^{s_d} \left[\mathcal{G}^{+}(x,z,s')e^{i\alpha_0(x-x')}\Phi^{+}(s')e^{i\alpha_0 x'} + \mathcal{N}^{+}(x,z,s')e^{i\alpha_0(x-x')}\Psi^{+}(s')e^{i\alpha_0 x'} \right] ds'. \tag{4.155}$$

*The result is that the expression of the **total** field in V^{+} can be obtained from its limit value Ψ^{+} on S and from its normal derivative Φ^{+}, by adding the incident field to the integral expression deduced from the Green theorem.*

The generalized compatibility equation is derived from the compatibility equations for F^{+} in V^{+} and for F^{i} in V^{-}**:**

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^{+}(s,s')\phi^{+}(s') + \mathcal{N}^{+}(s,s')\psi^{+}(s') \right] ds' - \frac{\psi^{+}(s)}{2} = 0, \tag{4.156}$$

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^{+}(s,s')\phi^{i}(s') + \mathcal{N}^{+}(s,s')\psi^{i}(s') \right] ds' + \frac{\psi^{i}(s)}{2} = 0. \tag{4.157}$$

By adding these two equations, we deduce that:

$$\int_{s'=0}^{s_d} \left[\mathcal{G}^{+}(s,s')\Phi^{+}(s') + \mathcal{N}^{+}(s,s')\Psi^{+}(s') \right] ds' + \psi^{i} = \frac{\Psi^{+}(s)}{2}. \tag{4.158}$$

The remarkable result is that the compatibility equation given in the preceding section can be extended to the total field: the left-hand side of eq. (4.158) represents the expression of the total field on S and the right-hand, one half of its limit on S. Thus the generalized compatibility condition can be stated in the following way:

The compatibility condition for a field in V^{+} including incident and/or diffracted waves and satisfying the two conditions:

- it is pseudo-periodic along the x axis:

$$F(x+d,z) = F(x,z)\exp(i\alpha_0 d), \tag{4.159}$$

- it satisfies a Helmholtz equation:

$$\nabla^2 F^{\pm} + k^2 (n^{\pm})^2 F^{\pm} = 0 \quad \text{in } V^{\pm}, \tag{4.160}$$

can be stated in the following way: *The value on S of the **total** field, obtained by adding the incident wave to the integral expression deduced from the second Green theorem (but in which the limits are those of the total field) is equal to the half of its limit value on S.*

A similar generalized compatibility condition can be obtained for a total field in V^{-} when an incident wave propagating upward in V^{-} (supposed to contain a lossless dielectric) illuminates the grating surface, but this case is not worth in the frame of this chapter.

4.A.6. Normal derivative of a field continuous on S.

The calculation of the normal derivative of $F^{\pm}$ on the grating surface in the general case is difficult. However, this aim can be reached at least in one case: when it is possible to define both F^+ and F^- which satisfy three conditions:

- F is continuous across S, or equivalently $\psi^+(s) = \psi^-(s)$,
- F^+ and F^- satisfy the same Helmholtz equation, with constant $k^2(n^+)^2$, or equivalently $n^+ = n^-$.
- F satisfies a radiation condition at infinity.

Due to the second condition, it seems that this case does not make sense: if $n^+ = n^-$, one cannot expect any scattering phenomenon. However, the study of this purely mathematical problem is crucial, for example in the study of perfectly conducting gratings.

First, it is worth noting that, thanks to the first condition, the gradient of F can be calculated without any use of distributions, which is not the case if F is discontinuous on S. Since $\gamma_m^- = \gamma_m^+$, $\mathcal{G}^-(x,z,s') = \mathcal{G}^+(x,z,s')$ and $\mathcal{N}^-(x,z,s') = \mathcal{N}^+(x,z,s')$ it can be deduced from eq. (4.139) that:

$$F^+(x,z) = \int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0 x} \phi^+(s') + \mathcal{N}^+(x,z,s') e^{i\alpha_0 x} \psi^+(s') \right] ds', \qquad (4.161)$$

$$F^-(x,z) = -\int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0 x} \phi^-(s') + \mathcal{N}^+(x,z,s') e^{i\alpha_0 x} \psi^-(s') \right] ds'. \qquad (4.162)$$

We have seen in this appendix that, if $\phi^+(s')$ and $\psi^+(s')$ are compatible, the expression of $F^+(x,z)$ given by eq. (4.161) vanishes in V^-. The same property holds for the expression of $F^-(x,z)$ given by eq. (4.162), which vanishes in V^+ if $\phi^-(s')$ and $\psi^-(s')$ are compatible. Bearing in mind that $\psi^+(s) = \psi^-(s)$, the expression of F in the entire space is given by adding the right-hand sides of eqs. (4.161) and (4.162):

$$F(x,z) = F^+(x,z) + F^-(x,z) = \int_{s'=0}^{s_d} \mathcal{G}^+(x,z,s') e^{i\alpha_0 x} \left(\phi^+(s') - \phi^-(s') \right) ds'. \qquad (4.163)$$

Thanks to the continuity of F on S, the expression of its value on the profile does not make problem and is given by:

$$F(s) = \int_{s'=0}^{s_d} \mathcal{G}^+(s,s') e^{i\alpha_0 x} \left(\phi^+(s') - \phi^-(s') \right) ds'. \qquad (4.164)$$

In order to obtain the normal derivative of F, let us calculate its gradient:

$$\nabla F\ (x,z) = \int_{s'=0}^{s_d} \nabla_{(x,z)}\left[\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}\right]\left(\phi^{+}(s')-\phi^{-}(s')\right)ds', \tag{4.165}$$

$$\nabla\left[\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}\right] = \frac{1}{2d}\begin{pmatrix} \sum\limits_{m=-\infty,+\infty} \frac{\alpha_m}{\gamma_m^{+}} e^{i\alpha_0 x} e^{imK(x-x')+i\gamma_m^{+}|z-z'|} \\ \sum\limits_{m=-\infty,+\infty} e^{i\alpha_0 x} e^{imK(x-x')+i\gamma_m^{+}|z-z'|} \end{pmatrix}. \tag{4.166}$$

The components of the normal $\vec{N}_S$ are given by:

$$\vec{N}_S = \begin{pmatrix} -\frac{dy}{ds} \\ \frac{dx}{ds} \end{pmatrix}, \tag{4.167}$$

and thus:

$$\left[\frac{d\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}}{dN_S}\right]^{\pm} = \frac{1}{2d}\lim_{\pm}\left\{\sum_{m=-\infty,+\infty}\left[\operatorname{sgn}(z-z')\frac{dx}{ds}-\frac{\alpha_m}{\gamma_m^{+}}\frac{dy}{ds}\right]e^{i\alpha_0 x}e^{imK(x-x')+i\gamma_m^{+}|z-z'|}\right\}. \tag{4.168}$$

Using eqs. (4.165) and (4.168) yields:

$$\frac{dF^{\pm}}{dN_S} = \int_{s'=0}^{s_d}\left[\frac{d\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}}{dN_S}\right]^{\pm}\left(\phi^{+}(s')-\phi^{-}(s')\right)ds'. \tag{4.169}$$

In order to eliminate the use of limits in the expression of $\left[\frac{d\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}}{dN_S}\right]^{\pm}$ given by eq. (4.168), it can be remembered that, by definition,

$$\frac{dF^{+}}{dN_S}-\frac{dF^{-}}{dN_S} = \left[\phi^{+}(s)-\phi^{-}(s)\right]e^{i\alpha_0 x}. \tag{4.170}$$

Moreover, it is to be noticed that, when z is constant, the components of $\nabla\left[\mathcal{G}^{+}(x,z,s')e^{i\alpha_0 x}\right]$ given by eq. (4.166) are Fourier series in x. Using again the property of discontinuous Fourier series on the discontinuity, it can be derived that:

$$\frac{dF^+}{dN_S}+\frac{dF^-}{dN_S}=$$
$$\frac{1}{d}\int_{s'=0}^{s_d}\sum_{m=-\infty}^{+\infty}\left[\operatorname{sgn}(z-z')\frac{dx}{ds}-\frac{\alpha_m}{\gamma_m^+}\frac{dy}{ds}\right]e^{i\alpha_0 x}e^{imK(x-x')+i\gamma_m^+|z-z'|}\left[\phi^+(s')-\phi^-(s')\right]ds'. \tag{4.171}$$

From equations (4.170) and (4.171), we deduce:

$$\frac{dF^\pm}{dN_S}=\pm\frac{\phi^+(s)-\phi^-(s)}{2}e^{i\alpha_0 x}+\int_{s'=0}^{s_d}\mathcal{K}(s,s')e^{i\alpha_0 x}\left[\phi^+(s')-\phi^-(s')\right]ds', \tag{4.172}$$

with:

$$\mathcal{K}(s,s')=\frac{1}{2d}\sum_{m=-\infty,+\infty}\left[\operatorname{sgn}(z-z')\frac{dx}{ds}-\frac{\alpha_m}{\gamma_m^+}\frac{dy}{ds}\right]e^{imK(x-x')+i\gamma_m^+|z-z'|}. \tag{4.173}$$

It is interesting to notice that the expression of function $\mathcal{K}(s,s')$ is very close to that of $\mathcal{N}(s,s')$, the only difference being that $\frac{dx'}{ds'}$ and $\frac{dy'}{ds'}$ are replaced by $\frac{dx}{ds}$ and $\frac{dy}{ds}$.

Like $\mathcal{N}^\pm(s,s')$, $\mathcal{K}(s,s')$ is continuous and its limit when $s'\to s$ can be expressed in closed form [44,7], as stated in section 6.3.

4.A.7. Limit values of a field with continuous normal derivative on S.

We consider a function F satisfying the following conditions

- F has the same normal derivative on both sides of S, or equivalently $\phi^+(s)=\phi^-(s)$,
- F^+ and F^- satisfy the same Helmholtz equation, or equivalently $n^+=n^-$.
- F satisfies a radiation condition at infinity.

The aim of this section is to calculate the limits of F on both parts of S.

From the secong Green theorem (eq. (4.136)), F can be expressed from the values of $\phi^+(s)$, $\phi^-(s)$, $\psi^+(s)$ and $\psi^-(s)$:

$$\begin{aligned}F^+(x,z)=\int_{s'=0}^{s_d}\Big[&\mathcal{G}^+(x,z,s')e^{i\alpha_0(x-x')}\phi^+(s')e^{i\alpha_0 x'}\\&+\mathcal{N}^+(x,z,s')e^{i\alpha_0(x-x')}\psi^+(s')e^{i\alpha_0 x'}\Big]ds',\end{aligned} \tag{4.174}$$

and, bearing in mind that $\phi^-(s)=\phi^+(s)$ and that $n^-=n^+$, $\mathcal{G}^-(x,z,s')=\mathcal{G}^+(x,z,s')$ and $\mathcal{N}^-(x,z,s')=\mathcal{N}^+(x,z,s')$:

$$F^-(x,z) = -\int_{s'=0}^{s_d} \left[\mathcal{G}^+(x,z,s') e^{i\alpha_0 (x-x')} \phi^+(s') e^{i\alpha_0 x'} + \mathcal{N}^+(x,z,s') e^{i\alpha_0 (x-x')} \psi^-(s') e^{i\alpha_0 x'} \right] ds'. \tag{4.175}$$

It has been shown in section 4.A.4 that, if $\phi^+(s')$ and $\psi^+(s')$ are compatible, the expression of $F^+(x,z)$ given by eq. (4.174) vanishes in V^-. The same property holds for the expression of $F^-(x,z)$ given by eq. (4.175), which vanishes in V^+, thus we can write that, if the compatibility equations are satisfied:

$$F\ (x,z) = F^+(x,z) + F^-(x,z) = \int_{s'=0}^{s_d} \mathcal{N}^+(x,z,s') e^{i\alpha_0 x} \Psi(s') ds', \tag{4.176}$$

or

$$U(x,z) = F(x,z) e^{-i\alpha_0 x} = \int_{s'=0}^{s_d} \mathcal{N}^+(x,z,s') \Psi(s') ds', \tag{4.177}$$

with:

$$\Psi(s') = \psi^+(s') - \psi^-(s'). \tag{4.178}$$

In order to express the limits $\lim\left\{U^+(x,z)\right\}$ or $\lim\left\{U^-(x,z)\right\}$ on both parts of S,. we can use a first equation:

$$\lim_+ \left\{U^+(x,z)\right\} - \lim_- \left\{U^-(x,z)\right\} = \Psi(s). \tag{4.179}$$

To find a second equation, we can consider eqs. (4.176) and (4.177) which gives the expression of $U(x,z)$. If z is fixed, the expression of $U(x,z)$ is a Fourier series in x, which is discontinuous on S. The value on S of this Fourier series is the average value of the limits on both sides of S, thus:

$$\lim_+ \left\{U^+(x,z)\right\} + \lim_- \left\{U^-(x,z)\right\} = 2\int_{s'=0}^{s_d} \mathcal{N}^+(s,s') \Psi(s') ds'. \tag{4.180}$$

From eqs. (4.179) and (4.180), we deduce the two limits:

$$\lim_\pm \left\{U(x,y)\right\} = \pm\frac{\Psi(s)}{2} + \int_{s'=0}^{s_d} \mathcal{N}^+(s,s') \Psi(s') ds'. \tag{4.181}$$

4.A.8. Calculation of the amplitudes of the plane wave expansions at infinity.

For $z \to \pm\infty$, the expression of $F^\pm$ given by eq. (4.136) can be simplified since $\mathrm{sgn}(z-z') = \pm 1$ and $|z-z'| = \pm(z-z')$, in such a way that the expression of F at infinity becomes a sum of plane waves:

$$F^+(x,y) = \sum_{m=-\infty}^{\infty} r_m \exp\left(i\alpha_m x + i\gamma_m^+ z\right) \quad \text{if } z > z_0\,, \tag{4.182}$$

$$F^-(x,y) = \sum_{m=-\infty}^{\infty} t_m \exp\left(i\alpha_m x - i\gamma_m^- z\right) \quad \text{if } z < 0\,, \tag{4.183}$$

$$r_m = \frac{1}{2d}\int_{s=0}^{s_d} e^{-imKx(s)-i\gamma_m^+ z(s)}\left[\frac{-i\phi^+(s)}{\gamma_m^+} + \left(\frac{dx(s)}{ds} - \frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\psi^+(s)\right]ds\,, \tag{4.184}$$

$$t_m = \frac{1}{2d}\int_{s=0}^{s_d} e^{-imKx(s)+i\gamma_m^- z(s)}\left[\frac{i\phi^-(s)}{\gamma_m^-} + \left(\frac{dx(s)}{ds} + \frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\psi^-(s)\right]ds\,, \tag{4.185}$$

with z_0 being the ordinate of the top of the grating profile. It must be noticed that a finite number of orders m, called propagating orders, are non-evanescent and propagate at infinity. They correspond to real values of γ_m^+ (for reflected orders) or γ_m^- (for transmitted orders, if the optical index n^- is real only).

Equation (4.184) can easily be generalized to the case in which we know the limit value of the total field on S (including incident waves) and its normal derivative. It suffices to analyze the behaviour at infinity of eq. (4.155) instead of eq. (4.136). The result is that it suffices to replace the values $\phi^+(s)$ and $\psi^+(s)$ relative to the scattered field by the values $\Phi^+(s)$ and $\Psi^+(s)$ relative to the total field:

$$r_m = \frac{1}{2d}\int_{s=0}^{s_d} e^{-imKx(s)-i\gamma_m^+ z(s)}\left[\frac{-i\Phi^+(s)}{\gamma_m^+} + \left(\frac{dx(s)}{ds} - \frac{\alpha_m}{\gamma_m^+}\frac{dz(s)}{ds}\right)\Psi^+(s)\right]ds\,. \tag{4.186}$$

Diffraction efficiencies ρ_m in the reflected orders propagating above the grating can be obtained by using the Poynting theorem on segments of one period parallel to the x axis:

$$\rho_m = \frac{\gamma_m^+}{\gamma_0^+}|r_m|^2\,. \tag{4.187}$$

When the grating is made of a lossless dielectric, the transmitted efficiencies τ_m are given by:

$$\tau_m = \frac{q^-}{q^+}\frac{\gamma_m^+}{\gamma_0^+}|t_m|^2\,. \tag{4.188}$$

Appendix 4.B. Integral method leading to a single integral equation for bare, metallic or dielectric grating

Historically, the formalism presented in this Appendix was the first one to lead to a single integral equation for a dielectric or metallic grating. The steps of the method are summarized in figure 4.14.

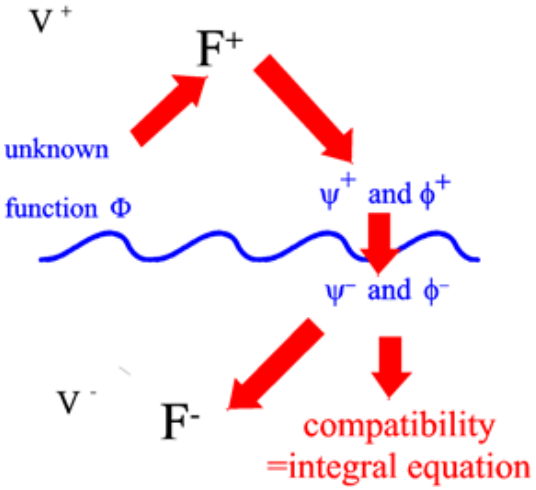


Figure 4.14. Steps of the integral formalism leading to a single equation

4.B.1. Definition of the unknown function

The single unknown function Φ is defined from a function $\tilde{\Gamma}=\begin{cases}\tilde{\Gamma}^{+} \text{ in } V^{+},\\ \tilde{\Gamma}^{-} \text{ in } V^{-},\end{cases}$ satisfying the following conditions:

- it satisfies the same Helmholtz equation in the entire space, except may be on the profile S,:

$$\nabla^2\tilde{\Gamma}+k^2\left(n^{+}\right)^2\tilde{\Gamma}=0, \tag{4.189}$$

- it has the same pseudo-periodicity as the actual solution of the grating problem:

$$\tilde{\Gamma}(x+d,z)=\tilde{\Gamma}(x,z)e^{i\alpha_0 d}, \tag{4.190}$$

- it identifies to the actual physical solution of the grating problem in V^{+}:

$$\tilde{\Gamma}^{+}\equiv F^{+}, \tag{4.191}$$

- it is continuous across S,
- it satisfies a radiation condition for $y\to\pm\infty$.

We denote by $\psi'^{\pm}e^{i\alpha_0 x'}$ and $\phi'^{\pm}e^{i\alpha_0 x'}$ the limit values of $\tilde{\Gamma}$ and of its normal derivative $\frac{d\tilde{\Gamma}}{dN_S}$ on S, bearing in mind that by definition, $\psi'^{+}\equiv\psi^{+}$ and $\phi'^{+}\equiv\phi^{+}$.

The question which arises is to know if $\tilde{\Gamma}^-$ is well defined. The continuity of $\tilde{\Gamma}$ across S imposes the value of $\tilde{\Gamma}^-$ on S: $\psi'^- \equiv \psi'^+ \equiv \psi^+$. In addition, $\tilde{\Gamma}^-$ satisfies a Helmholtz equation and a radiation condition at infinity. It is worth noting that, in contrast with the function Γ introduced in Appendix 4.A, here the function $\tilde{\Gamma}$ does not include the incident field and thus has no physical meaning below the profile. As mentioned in Appendix 4.A, the solution of this boundary value problem (since we impose that $\tilde{\Gamma}^- = \tilde{\Gamma}^+$) exists and is unique, thus $\tilde{\Gamma}$ is correctly defined in the entire space. The unknown function Φ of the integral equation is defined as the jump of the normal derivative of $\tilde{\Gamma}$ across S, more precisely:

$$\Phi = \phi'^+ - \phi'^- . \tag{4.192}$$

As regards the physical interpretation of Φ, it can be shown easily that $\Phi\, e^{i\alpha_0 x}$ is the surface current density which, placed on S, generates in V^+ the actual scattered field. In other words, we have considered a fictitious structure consisting of an infinitely thin, perfectly conducting metallic sheet supporting a surface current density $j(s)\hat{y}$ placed on the grating surface, this surface separating two media having identical optical properties (refractive index n^+). The unknown Φ is proportional to j.

4.B.2. Expression of the scattered field, its limit on S and its normal derivative from Φ.

It must be noticed that the function $\tilde{\Gamma}$ satisfies all the conditions of the function F of section 4.A.5. Since $\tilde{\Gamma}$ is continuous on S, the calculation of $\tilde{\Gamma}$ from Φ can be achieved using eqs. (4.163) and (4.137):

$$\tilde{\Gamma}(x,z) = \int_{s'=0}^{s_d} \mathcal{G}^+(x,z,s') e^{i\alpha_0 x} \Phi(s') ds', \tag{4.193}$$

$$\mathcal{G}^+(x,y,s') = \frac{1}{2id} \sum_{m=-\infty}^{\infty} \frac{1}{\gamma_m^+} \exp\left[imK(x-x') + i\gamma_m^+ |z-z'| \right]. \tag{4.194}$$

The value of the limit $\psi'^+(s) e^{i\alpha_0 x}$ of $\tilde{\Gamma}^+(x,z)$ on S does not make problem, thanks to its continuity:

$$\psi'^+(s) = \int_{s'=0}^{s_d} \mathcal{G}^+(s,s') \Phi(s') ds', \tag{4.195}$$

with $\mathcal{G}^+(s,s')$ the value on S of $\mathcal{G}^+(x,y,s')$, given by eq. (4.147):

$$\mathcal{G}^+(s,s') = \frac{1}{2id} \sum_{m=-\infty}^{\infty} \frac{1}{\gamma_m^+} e^{imK(x(s)-x'(s'))+i\gamma_m^+ |z(s)-z'(s')|}, \tag{4.196}$$

and finally its normal derivative can be derived from eq. (4.172) and (4.173):

$$\phi^+(s) = \frac{\Phi(s)}{2} + \int_{s'=0}^{s_d} \mathcal{K}(s,s') \Phi(s') ds', \tag{4.197}$$

$$\mathcal{K}(s,s') = \frac{1}{2d}\sum_{m=-\infty,+\infty}\left[\operatorname{sgn}(z-z')\frac{dx}{ds} - \frac{\alpha_m}{\gamma_m^+}\frac{dy}{ds}\right] e^{imK(x-x')+i\gamma_m^+|z-z'|}. \qquad (4.198)$$

It is worth noting that the values of the limits of the field and its normal derivative on S that are the two unknown functions in the classical integral theory (section 2), are now expressed from the single function $\Phi(s)$. This is not surprising since the limits on S of the function $\tilde{\Gamma}$ given by eq. (4.193) satisfy automatically a relation of compatibility, whatever the function $\Phi(s')$ introduced in the integral may be: it is the field generated by a surface current on S. As a consequence, we have not to include in the theory a relation of compatibility in V^+, which was the first integral equation in the classical formalism.

4.B.3. Integral equation

The single integral equation will be obtained by writing the relation of compatibility in V^-, considered now to be filled by the actual grating material, i.e. a material of index n^-. With this aim, we calculate the limits in V^- of the field and its normal derivative on S. inserting in the continuity conditions of the field given by eqs (4.8) and (4.10) the the limit values given by eqs (4.195) and (4.197):

$$\psi^-(s) = \psi^i(s) + \int_{s'=0}^{s_d} \mathcal{G}^+(s,s')\Phi(s')ds', \qquad (4.199)$$

$$\phi^-(s) = \frac{q^+}{q^-}\left[\frac{\Phi(s)}{2} + \phi^i(s) + \int_{s'=0}^{s_d} \mathcal{K}(s,s')\Phi(s')ds'\right], \qquad (4.200)$$

with $q^\pm$ given by eq. (4.12).

The field in V^- can be deduced from eqs. (4.199) and (4.200) using eq. (4.136):

$$F^-(x,z) = -e^{i\alpha_0 x}\int_{s'=0}^{s_d}\left[\mathcal{G}^-(x,z,s')\phi^-(s') + \mathcal{N}^-(x,z,s')\psi^-(s')\right]ds', \qquad (4.201)$$

and the equation of compatibility is given by eq. (4.146):

$$\begin{aligned}&\int_{s'=0}^{s_d}\left\{\frac{q^+}{q^-}\mathcal{G}^-(s,s')\left[\frac{\Phi(s')}{2} + \phi^i(s') + \int_{s''=0}^{s_d}\mathcal{K}(s',s'')\Phi(s'')ds''\right]\right.\\ &\left.+\mathcal{N}^-(s,s')\left[\psi^i(s') + \int_{s''=0}^{s_d}\mathcal{G}^+(s',s'')\Phi(s'')ds''\right]\right\}ds' + \frac{\psi^i(s)}{2} + \frac{1}{2}\int_{s'=0}^{s_d}\mathcal{G}^+(s,s')\Phi(s')ds' = 0,\end{aligned} \qquad (4.202)$$

which yields, after simplification:

$$\left[\left(\mathcal{N}^- + \frac{\mathbb{I}}{2}\right)\mathcal{G}^+ + \frac{q^+}{q^-}\mathcal{G}^-\left(\mathcal{K} + \frac{\mathbb{I}}{2}\right)\right]\Phi = -\left(\mathcal{N}^- + \frac{\mathbb{I}}{2}\right)\psi^i(s) + \frac{q^+}{q^-}\mathcal{G}^-\phi^i, \qquad (4.203)$$

with the symbol $\mathcal{O}\eta$ denoting the function $\int_{s'=0}^{s_d} \mathcal{O}(s,s')\eta(s')ds'$ in operator notation.

For $z \to +\infty$, the expression of $\tilde{\Gamma}^+ \equiv F^+$ given by eq. (4.193) can be simplified since $\mathrm{sgn}(z-z') = \pm 1$ and $|z-z'| = \pm(z-z')$, in such a way that the expression of F^+ at infinity becomes a sum of plane waves:

$$F^+(x,y) = \sum_{m=-\infty}^{\infty} r_m \exp\left(i\alpha_m x + i\gamma_m^+ z\right) \quad \text{if } z > z_0\,, \tag{4.204}$$

with amplitudes given by:

$$r_m = \frac{1}{2id\gamma_m^+} \int_{s=0}^{s_d} e^{-imKx(s) - i\gamma_m^+ z(s)} \Phi(s)\, ds\,. \tag{4.205}$$

The diffraction efficiencies of the reflected waves are then deduced by:

$$\rho_m = \frac{\gamma_m^+}{\gamma_0^+} |r_m|^2\,. \tag{4.206}$$

Similarly, it can be derived from eq. (4.201) that the transmitted field can be represented by a sum of plane waves below the profile:

$$F^-(x,y) = \sum_{m=-\infty}^{\infty} t_m \exp\left(i\alpha_m x + i\gamma_m^- z\right) \quad \text{if } z < 0\,. \tag{4.207}$$

The amplitudes of the transmitted plane waves are derived from eq. (4.185) after calculating the functions ϕ^- and ψ^- from eqs. (4.199) and (4.200):

$$t_m = \frac{1}{2d} \int_{s=0}^{s_d} e^{-imKx(s) + i\gamma_m^- z(s)} \left[\frac{i\phi^-(s)}{\gamma_m^-} + \left(\frac{dx(s)}{ds} + \frac{\alpha_m}{\gamma_m^+} \frac{dz(s)}{ds} \right) \psi^-(s) \right] ds\,. \tag{4.208}$$

It is interesting to notice that this method has been extended to other problems (including 3D problems of scattering) by many specialists of theoretical physics and applied mathematics [49-52].

References:

[1] R. Courant, D. Hilbert: *Methods of Mathematical Physics*, v. 1 (Interscience, New Your, 1965), ch.3, pp. 112-163.
[2] P. M. Morse, H. Feshbach : *Methods of Theoretical Physics*, Part 1 (Mc Graw-Hill, New York, 1953), ch.8, pp.896-997
[3] L. Schwartz: *Méthodes Mathématiques pour les Sciences Physiques* (Hermann, Paris, 1965)
[4] L. Schwartz : *Théorie des Distributions* (Hermann, Paris, 1966)
[5] L. Schwartz, *Mathematics for the physical sciences*. *(*Hermann; Addison-Wesley Publishing Co., Reading, 1966)
[6] R. Petit, ed., *Electromagnetic Theory of Gratings* (Springer, Berlin, 1980)
[7] D. Maystre: Thèse d'Etat, Marseille AO 9545 (1974)
[8] R. Petit, M. Cadilhac: C. R. Acad. Sci. Paris, **259**, 2077 (1964)
[9] A. Wirgin: Rev. Opt., **9**, 449 (1964)
[10] J. L. Uretsky: Ann. Phys. **33**, 400 (1965)
[11] R. Petit: C. R. Acad Sci. Paris, **260**, 4454 (1965)
[12] R. Petit: Rev. Opt. **45**, 249 (1966)
[13] J. Pavageau, R. Eido, H. Kobeissé: C. R. Acad. Sci. Paris, **264**, 424 (1967)
[14] A. Wirgin: Rev. Cethedec, **5**, 131 (1968)
[15] A. Neureuther, K. Zaki: Alta Freq. **38**, 282 (1969)
[16] P. M. Van den Berg : Thesis, Delft, the Netherlands (1971)
[17] D. Maystre: Opt. Commun., **6**, 50 (1972)
[18] D. Maystre: Opt. Commun., **8**, 216 (1973)
[19] R. Petit, D. Maystre, M. Nevière: Space Optics Proc. 9th Congr. I.C.O., **667** (1972)
[20] D. Maystre : J. Opt. Soc. Am. **68**, 490 (1978)
[21] D. Maystre: Opt. Commun. **26**, 127 (1978)
[22] L. C. Botten: Opt. Acta **25**, 481 (1978)
[23] L. C. Botten: Ph.D; Thesis, Tasmania, Hobart (1978)
[24] G. H. Spencer, M. V. Murty : J. Opt. Soc. Am. **52**, 672 (1962)
[25] W. Werner: Thesis (1970).
[26] D. Maystre, R. Petit: Opt. Commun. **4**, 97 (1971).
[27] R. Petit, D. Maystre : Rev. Phys. Appl. **7**, 427 (1972).
[28] D. Maystre. Rigorous vector theories of diffraction gratings. Progress in Optics, **21**, 1 (1984).
[29] A. W. Maue: Z. Phys. **126**, 601 (1949)
[30] M. Nevière, M. Cadilhac : Opt. Commun. **4**, 13 (1971)
[31] D. Maystre, R. Petit : Opt. Commun. **5**, 35 (1972)
[32] R. Petit, D. Maystre : Rev. Phys. Appl. **7**, 427 (1972)
[33] D. Maystre, R. Petit : J. Spectr. Soc. Jpn. **23** suppl. 61 (1974)
[34] G. Schmidt and B.H. Kleemann, Journal of Modern Optics, **58**,. 407 (2011).
[35] D. Maystre, R. Petit : Opt. Commun. **2**, 309 (1970).
[36] D. Maystre, R. Petit: C. R. Acad. Sci. Paris **271**, 400 (1970).
[37] D. Maystre, R. Petit : Opt. Commun. **4**, 25 (1971).
[38] R. C. McPhedran : Ph. D. Thesis, Tasmania, Hobart (1973).
[39] E. Popov, B. Bozhkov, D. Maystre, and J. Hoose, Applied Optics, **38**, 47 (1999).
[40] G. Dumery, P. Filippi : C. R. Acad. Sci Paris **270**, 137 (1970)
[41] H.Kalhor, A. Neureuther : J. Opt. Soc. Am. **61**, 43 (1971)
[42] P. M. van der Berg: Appl. Sci. Res. **24**, 261 (1971)
[43] R. Green: IEEE Trans. MTT-**18**, 313 (1970)

[44] J. Pavageau, J. Bousquet: Opt. Acta **17**, 469 (1970)
[45] M. Abramowitz and I.A. Stegun, Handbook of mathematical functions(Dover Publications, New-York, 1964).
[46] J. Meixner: IEEE Trans. AP-**20**, 442 (1972).
[47] L. Li and J. Chandezon, J. Opt. Soc. Am. **13**, 2247 (1996)
[48] A. Marechal and G. W. Stroke, C. R. Acad. Sci., Paris, **249**, 2042 (1959).
[49] R.E. Kleinman and P.A. Martin, SIAM J. Appl. Math. , **48,** 307 (1988).
[50] E. Marx, J. Math. Phys. **23**, 1057 (1982).
[51] A. W. Glisson, IEEE Trans. Antennas Propagation, vol. **32**, 173 (1984).
[52] M.S. Yeung, IEEE Trans. Antennas and Propagation, **47**, 1615 (1999).

Chapter 5:

Finite element Method

Guillaume Demésy,

Frédéric Zolla,

André Nicolet, and

Benjamin Vial

Table of Contents:

Chapter 5

Finite Element Method

Guillaume Demésy, Frédéric Zolla, André Nicolet, and Benjamin Vial

Aix-Marseille Université, École Centrale Marseille, Institut Fresnel,
13397 Marseille Cedex 20, France
guillaume.demesy@fresnel.fr

5.1 Introduction

Finite element methods (FEM) represent a very general set of techniques to approximate solutions of partial derivative equations. Their main advantage lies in their ability to handle arbitrary geometries via unstructured meshes of the domain of interest: The discretization of oblic geometry edges is natively built in. Finite Element Methods have been widely developed in many areas of physics and engineering: mechanics, thermodynamics...

But until the early 80's, two major drawbacks prevented them from being used in electromagnetic problems. On the one hand, existing nodal element basis did not satisfy the physical (dis)continuity of the vector fields components and lead to spurious solutions [1]. On the other hand, there was no proper way to truncate unbounded regions in open wave problems.

These two major limitations were both overcome in the early 80's: Vector elements have been developed by Nédélec [2, 3], and Perfectly Matched Layers (PMLs) were discovered by Bérenger [4]. Since then, it has been shown that PMLs could be described in the general framework of transformation optics [5, 6, 7, 8].

All the mathematical and computational ingredients now exist and the goal of this chapter is to show how to combine them to implement a general 3D numerical scheme adapted to gratings using Finite Elements. In fact, we are now facing the physical difficulties inherent to the infinite spatial characteristics of the grating problem, whereas the computation domain has to be bounded in practice: (i) Both the superstrate and the substrate are infinite regions, (ii) there is an infinite number of periods and, last but not least, (iii) the sources of the incident field (a plane wave) are located in the superstrate at an infinite distance from the grating.

In this chapter, the infinite extension of the superstrate and substrate is addressed using cartesian PMLs. In the framework of transformation optics, we demonstrate that Bérenger's original PMLs can be extended to the challenging numerical cases of grazing incidence in order to deal with extreme oblic incidences or configurations near Wood's anomalies. The second issue of infinite number of period can be addressed via Bloch conditions. Finally, we are dealing with the distant plane wave sources through an equivalence of the diffraction problem with a radiation one whose sources are localized inside the diffractive element itself. The unknown field to be approximated using Finite Elements is a *radiated field* with sources *inside* the computation box and allows to retrieve easily the *total field* with the plane wave source.

In a first section, we derive and implement this approach in the so-called 2D non-conical, or scalar, case. We are dealing with the infinite issues rigorously in both TE and TM polarization cases. It results in a radiation problem with sources localized in the diffractive element itself. We mathematically split the whole problem into two parts. The first one consists in the classical calculation of the *total field* solution of a simple interface. The second one amounts to looking

for a *radiated field* with sources confined within the diffractive obstacles and deduced from the first elementary problem. From this viewpoint, the later *radiated field* can be interpreted as an *exact perturbation* of the *total field*. We show that our approach allows to tackle some kind of anisotropy without increasing the computational time or resource. Through a battery of examples, we illustrate its independence towards the geometry of the diffractive pattern. Finally, we present an Adaptative PML able to tackle grazing incidences or configurations near Wood's anomaly.

In a second section, we extend this approach to the most general configuration of vector diffraction by crossed gratings embedded in arbitrary multilayered stack. The main advantage of this method is, again, its complete independence towards the shape of the diffractive element, whereas other methods often require heavy adjustments depending on whether the geometry of the groove region presents oblique edges. This approach combined with the use of second order edge elements allows us to retrieve the few numerical academic examples found in the literature with an excellent accuracy. Furthermore, we provide a new reference case combining major difficulties: A non trivial toroidal geometry together with strong losses and a high permittivity contrast. Finally, we discuss computation time and convergence as a function of the mesh refinement as well as the choice of the direct solver.

5.2 Scalar diffraction by arbitrary mono-dimensional gratings : a Finite Element formulation

5.2.1 Set up of the problem and notations

We denote by $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, the unit vectors of the axes of an orthogonal coordinate system $Oxyz$. We deal only with time-harmonic fields; consequently, the electric and magnetic fields are represented by the complex vector fields $\mathbf{E}$ and $\mathbf{H}$, with a time dependance in $\exp(-i\omega t)$.

Besides, in this chapter, we assume that the tensor fields of relative permittivity $\underline{\underline{\varepsilon}}$ and relative permeability $\underline{\underline{\mu}}$ can be written as follows:

$$\underline{\underline{\varepsilon}} = \begin{pmatrix} \varepsilon_{xx} & \bar{\varepsilon}_a & 0 \\ \varepsilon_a & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{pmatrix} \quad \text{and} \quad \underline{\underline{\mu}} = \begin{pmatrix} \mu_{xx} & \bar{\mu}_a & 0 \\ \mu_a & \mu_{yy} & 0 \\ 0 & 0 & \mu_{zz} \end{pmatrix}, \tag{5.1}$$

where $\varepsilon_{xx}, \varepsilon_a, \ldots \mu_{zz}$ are possibly complex valued functions of the two variables x and y and where $\bar{\varepsilon}_a$ (resp. $\bar{\mu}_a$) represents the conjugate complex of ε_a (resp. μ_a). *These kinds of materials are said to be z–anisotropic*. It is of importance to note that with such tensor fields, lossy materials can be studied (the lossless materials correspond to tensors with real diagonal terms represented by Hermitian matrices) and that the problem is invariant along the z–axis but the tensor fields can vary continuously (gradient index gratings) or discontinuously (step index gratings). Moreover we define $k_0 := \omega/c$.

The gratings that we are dealing with are made of three regions (See Fig. 5.1).

- *The superstratum* ($y > h^g$) which is supposed to be homogeneous, isotropic and lossless and characterized solely by its relative permittivity ε^+ and its relative permeability μ^+ and we denote $k^+ := k_0\sqrt{\varepsilon^+\mu^+}$
- *The substratum* ($y < 0$) which is supposed to be homogeneous and isotropic and therefore characterized by its relative permittivity ε^- and its relative permeability μ^- and we denote $k^- := k_0\sqrt{\varepsilon^-\mu^-}$

- *The groove region* ($0 < y < h^g$) which can be heterogeneous and z–anisotropic and thus characterized by the two tensor fields $\underline{\underline{\varepsilon}}^g(x,y)$ and $\underline{\underline{\mu}}^g(x,y)$. It is worth noting that the method does work irrespective of whether the tensor fields are piecewise constant. The groove periodicity along x–axis will be denoted d.

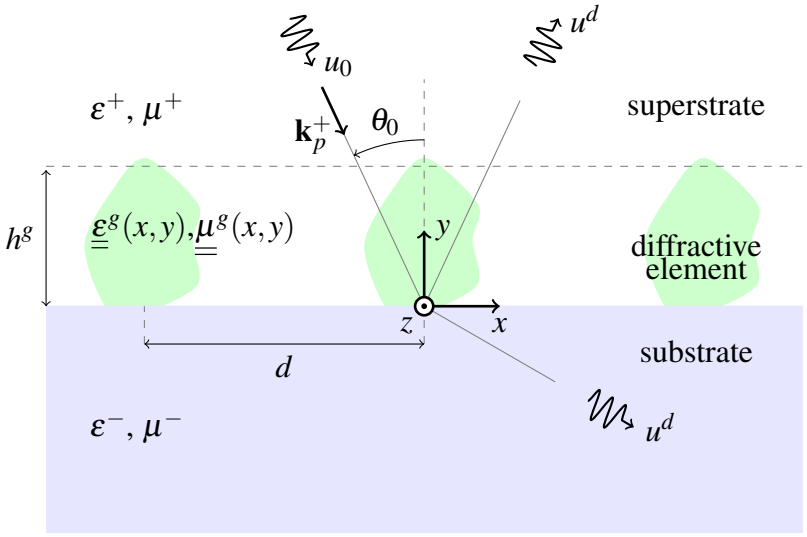


Fig. 5.1: *Sketch and notations of the grating studied in this section.*

This grating is illuminated by an incident plane wave of wave vector $\mathbf{k}_p^+ = \alpha\,\mathbf{x} - \beta^+\,\mathbf{y} = k^+(\sin\theta_0\mathbf{x} - \cos\theta_0\mathbf{y})$, whose electric field (TM case) (resp. magnetic field (TE case)) is linearly polarized along the z–axis:

$$\mathbf{E}_e^0 = \mathbf{A}_e^0 \exp(i\mathbf{k}_p^+ \cdot \boldsymbol{r})\,\mathbf{z} \quad (\text{resp. } \mathbf{H}_m^0 = \mathbf{A}_m^0 \exp(i\mathbf{k}_p^+ \cdot \boldsymbol{r})\,\mathbf{z})\,, \tag{5.2}$$

where $\mathbf{A}_e^0$ (resp. $\mathbf{A}_m^0$) is an arbitrary complex number and $\boldsymbol{r} = (x,y)^{\mathrm{T}}$. In this section, a plane wave is characterized by its wave-vector denoted $\mathbf{k}_{\{p,c\}}^{\{+,-\}}$. The subscript p (resp. c) stands for "propagative" (resp. "counter-propagative"). The superscript $+$ (resp. $-$) refers to the associated wavenumber k^+ (resp. k^-), and indicates that we are dealing with a plane wave propagating in the superstrate (resp. substrate).The magnetic (resp. electric) field derived from $\mathbf{E}_e^0$ (resp. $\mathbf{H}_m^0$) is denoted $\mathbf{H}_e^0$ (resp. $\mathbf{E}_m^0$) and the electromagnetic field associated with the incident field is therefore denoted $(\mathbf{E}^0, \mathbf{H}^0)$ which is equal to $(\mathbf{E}_e^0, \mathbf{H}_e^0)$ (resp. $(\mathbf{E}_m^0, \mathbf{H}_m^0)$).

The diffraction problem that we address consists in finding Maxwell equation solutions in harmonic regime *i.e.* the unique solution $(\mathbf{E}, \mathbf{H})$ of:

$$\begin{cases} \mathbf{curl}\,\mathbf{E} = i\,\omega\,\mu_0\,\underline{\underline{\mu}}\,\mathbf{H} & \text{(5.3a)} \\ \mathbf{curl}\,\mathbf{H} = -i\,\omega\,\varepsilon_0\,\underline{\underline{\varepsilon}}\,\mathbf{E} & \text{(5.3b)} \end{cases}$$

such that the diffracted field $(\mathbf{E}^{\mathrm{d}}, \mathbf{H}^{\mathrm{d}}) := (\mathbf{E} - \mathbf{E}_e^0, \mathbf{H} - \mathbf{H}_m^0)$ satisfies an *Outgoing Waves Condition* (O.W.C. [9]) and where $\mathbf{E}$ and $\mathbf{H}$ are quasi-periodic functions with respect to the x coordinate.

5.2.2 Theoretical developments of the method

5.2.2.1 Decoupling of fields and z–anisotropy

We assume that $\underline{\underline{\delta}}(x,y)$ is a z–anisotropic tensor field ($\delta_{xz}=\delta_{yz}=\delta_{zx}=\delta_{zy}=0$). Moreover, the left upper matrix extracted from $\underline{\underline{\delta}}$ is denoted $\tilde{\underline{\underline{\delta}}}$, namely:

$$\tilde{\underline{\underline{\delta}}} = \begin{pmatrix} \delta_{xx} & \bar{\delta}_a \\ \delta_a & \delta_{yy} \end{pmatrix} . \tag{5.4}$$

For z–anisotropic materials, in a non-conical case, the problem of diffraction can be split into two fundamental cases (TE case and TM case). This property results from the following equality which can be easily derived:

$$-\mathbf{curl}\left(\underline{\underline{\delta}}^{-1}\,\mathbf{curl}\,(u\,\mathbf{z})\right) = \mathrm{div}\left(\tilde{\underline{\underline{\delta}}}^{T}/\det(\tilde{\underline{\underline{\delta}}})\nabla u\right)\mathbf{z}\,, \tag{5.5}$$

where u is a function which does not depend on the z variable. Relying on the previous equality, it appears that the problem of diffraction in a non conical mounting amounts to looking for an electric (resp. magnetic) field which is polarized along the z–axis ; $\mathbf{E}=e(x,y)\,\mathbf{z}$ (resp. $\mathbf{H}=h(x,y)\,\mathbf{z}$). The functions e and h are therefore solutions of similar differential equations:

$$\mathscr{L}_{\underline{\underline{\xi}},\chi}(u) := \mathrm{div}\left(\underline{\underline{\xi}}\,\nabla u\right) + k_0^2\chi\,u = 0 \tag{5.6}$$

with

$$u=e, \quad \underline{\underline{\xi}} = \tilde{\underline{\underline{\mu}}}^{T}/\det(\tilde{\underline{\underline{\mu}}}), \quad \chi=\varepsilon_{zz}\,, \tag{5.7}$$

in the TM case and

$$u=h, \quad \underline{\underline{\xi}} = \tilde{\underline{\underline{\varepsilon}}}^{T}/\det(\tilde{\underline{\underline{\varepsilon}}}), \quad \chi=\mu_{zz}\,, \tag{5.8}$$

in the TE case.

5.2.2.2 Boiling down the diffraction problem to a radiation one

In its initial form, the diffraction problem summed up by Eq. (5.6) is not well suited to the Finite Element Method. In order to overcome this difficulty, we propose to split the unknown function u into a sum of two functions u_1 and u_2^d, the first term being known as a closed form and the latter being a solution of a problem of radiation *whose sources are localized within the obstacles*.

We have assumed that outside the groove region (cf. Fig. 5.1), the tensor field $\underline{\underline{\xi}}$ and the function χ are constant and equal respectively to $\underline{\underline{\xi}}^{-}$ and χ^{-} in the substratum ($y<0$) and equal respectively to $\underline{\underline{\xi}}^{+}$ and χ^{+} in the superstratum ($y>h^g$). Besides, for the sake of clarity, the superstratum is supposed to be made of an isotropic and lossless material and is therefore solely defined by its relative permittivity ε^{+} and its relative permeability μ^{+}, which leads to:

$$\underline{\underline{\xi}}^{+} = \frac{1}{\mu^{+}}\,\mathrm{Id}_2 \quad \text{and} \quad \chi^{+}=\varepsilon^{+} \quad \text{in TE case} \tag{5.9}$$

or

$$\underline{\underline{\xi}}^{+} = \frac{1}{\varepsilon^{+}}\,\mathrm{Id}_2 \quad \text{and} \quad \chi^{+}=\mu^{+} \quad \text{in TM case,} \tag{5.10}$$

where Id_2 is the 2×2 identity matrix. With such notations, $\underline{\underline{\xi}}$ and χ are therefore defined as follows:

$$\underline{\underline{\xi}}(x,y) := \begin{cases} \underline{\underline{\xi}}^+ & \text{for} \quad y > h^g \\ \underline{\underline{\xi}}^g(x,y) & \text{for} \quad h^g > y > 0 \\ \underline{\underline{\xi}}^- & \text{for} \quad y < 0 \end{cases}, \; \chi(x,y) := \begin{cases} \chi^+ & \text{for} \quad y > h^g \\ \chi^g(x,y) & \text{for} \quad h^g > y > 0 \\ \chi^- & \text{for} \quad y < 0\,. \end{cases} \tag{5.11}$$

It is now apropos to introduce an auxiliary tensor field $\underline{\underline{\xi}}_1$ and an auxiliary function χ_1:

$$\underline{\underline{\xi}}_1(x,y) := \begin{cases} \underline{\underline{\xi}}^+ & \text{for} \quad y > 0 \\ \underline{\underline{\xi}}^- & \text{for} \quad y < 0 \end{cases}, \; \chi_1(x,y) := \begin{cases} \chi^+ & \text{for} \quad y > 0 \\ \chi^- & \text{for} \quad y < 0\,, \end{cases} \tag{5.12}$$

these quantities corresponding, of course, to a simple plane interface. Besides, we introduce the constant tensor field $\underline{\underline{\xi}}_0$ which is equal to $\underline{\underline{\xi}}^+$ everywhere and a constant scalar field χ_0 which is equal to χ^+ everywhere. Finally, we denote u_0 the function which equals the incident field u^{inc} in the superstratum and vanishes elsewhere (see Fig. 5.1):

$$u_0(x,y) := \begin{cases} u^{\mathrm{inc}} & \text{for} \quad y > h^g \\ 0 & \text{for} \quad y < h^g \end{cases} \tag{5.13}$$

We are now in a position to define more precisely the diffraction problem that we are dealing with. The function u is the unique solution of:

$$\mathscr{L}_{\underline{\underline{\xi}},\chi}(u) = 0\,, \text{ such that } u^d := u - u_0 \text{ satisfies an O.W.C.} \tag{5.14}$$

In order to reduce this diffraction problem to a radiation problem, an intermediate function is necessary. This function, called u_1, is defined as the unique solution of the equation:

$$\mathscr{L}_{\underline{\underline{\xi}}_1,\chi_1}(u_1) = 0\,, \text{ such that } u_1^d := u_1 - u_0 \text{ satisfies an O.W.C.} \tag{5.15}$$

The function u_1 corresponds thus to *an annex problem* associated to a simple interface and can be solved in closed form and *from now on is considered as a known function*. As written above, we need the function u_2^d which is simply defined as the difference between u and u_1:

$$u_2^d := u - u_1 = u^d - u_1^d\,. \tag{5.16}$$

The presence of the superscript d is, of course, not irrelevant: As the difference of two diffracted fields, the O.W.C. of u_2^d is guaranteed (which is of prime importance when dealing with PML cf. 5.2.2.4). As a result, the Eq. (5.14) becomes:

$$\mathscr{L}_{\underline{\underline{\xi}},\chi}(u_2^d) = -\mathscr{L}_{\underline{\underline{\xi}},\chi}(u_1)\,, \tag{5.17}$$

where the right hand member is a scalar function which may be interpreted as a *known source term* $-\mathscr{S}_1(x,y)$ *and the support of this source is localized only within the groove region*. To prove it, all we have to do is to use Eq. (5.15):

$$\mathscr{S}_1 := \mathscr{L}_{\underline{\underline{\xi}},\chi}(u_1) = \mathscr{L}_{\underline{\underline{\xi}},\chi}(u_1) - \underbrace{\mathscr{L}_{\underline{\underline{\xi}}_1,\chi_1}(u_1)}_{=0} = \mathscr{L}_{\underline{\underline{\xi}}-\underline{\underline{\xi}}_1,\chi-\chi_1}(u_1)\,. \tag{5.18}$$

Now, let us point out that the tensor fields $\underline{\underline{\xi}}$ and $\underline{\underline{\xi}}_1$ are identical outside the groove region and the same holds for χ and χ_1. The support of $\mathscr{S}_1$ is thus localized within the groove region as expected. It remains to compute more explicitly the source term $\mathscr{S}_1$. Making use of the linearity of the operator $\mathscr{L}$ and the equality $u_1 = u_1^d + u_0$, the source term can be split into two terms

$$\mathscr{S}_1 = \mathscr{S}_1^0 + \mathscr{S}_1^d \,, \tag{5.19}$$

where

$$\mathscr{S}_1^0 = \mathscr{L}_{\underline{\underline{\xi}}-\underline{\underline{\xi}}_1,\chi-\chi_1}(u_0) \tag{5.20}$$

and

$$\mathscr{S}_1^d = \mathscr{L}_{\underline{\underline{\xi}}-\underline{\underline{\xi}}_1,\chi-\chi_1}(u_1^d) \,. \tag{5.21}$$

Now, bearing in mind that u_0 is nothing but a plane wave $u_0 = \exp(i\mathbf{k}_p^+ \cdot \mathbf{r})$ (with $\mathbf{k}_p^+ = \alpha\mathbf{x} - \beta^+\mathbf{y}$), it is sufficient to give $\nabla u_0 = i\mathbf{k}_p^+ u_0$ for the weak formulation associated with Eq. (5.17):

$$\mathscr{S}_1^0 = \left\{ i \operatorname{div}\left[\left(\underline{\underline{\xi}}^+ - \underline{\underline{\xi}} \right) \mathbf{k}_p^+ \exp(i\mathbf{k}_p^+ \cdot \mathbf{r}) \right] + k_0^2 \left(\chi^+ - \chi \right) \exp(i\mathbf{k}_p^+ \cdot \mathbf{r}) \right\} \,. \tag{5.22}$$

The same holds for the term associated with the diffracted field. Since, in the superstrate, we have of course $u_1^d = \rho \exp(i\mathbf{k}_c^+ \cdot \mathbf{r})$ with $\mathbf{k}_c^+ = \alpha\mathbf{x} + \beta^+\mathbf{y}$,

$$\mathscr{S}_1^d = \rho \left\{ i \operatorname{div}\left[\left(\underline{\underline{\xi}}^+ - \underline{\underline{\xi}} \right) \mathbf{k}_c^+ \exp(i\mathbf{k}_c^+ \cdot \mathbf{r}) \right] + k_0^2 \left(\chi^+ - \chi \right) \exp(i\mathbf{k}_c^+ \cdot \mathbf{r}) \right\} \,, \tag{5.23}$$

where ρ is simply the complex reflection coefficient associated with the simple interface:

$$\rho = \frac{p^+ - p^-}{p^+ - p^-} \text{ with } p^\pm = \begin{cases} \beta^\pm \text{ in the TM case} \\ \\ \frac{\beta^\pm}{\varepsilon^\pm} \text{ in the TE case} \end{cases} \tag{5.24}$$

5.2.2.3 Quasi-periodicity and weak formulation

The weak formulation follows the classical lines and is based on the construction of a weighted residual of Eq. (5.6), which is multiplied by the complex conjugate of a weight function u' and integrated by part to obtain:

$$\mathscr{R}_{\underline{\underline{\xi}},\chi}(u,u') = -\int_\Omega \left(\underline{\underline{\xi}} \nabla u \right) \cdot \nabla \overline{u'} + k_0^2 \chi \, u \, \overline{u'} \, \mathrm{d}\Omega + \int_{\partial\Omega} \overline{u'} \left(\underline{\underline{\xi}} \nabla u \right) \cdot \mathbf{n} \, \mathrm{d}S \tag{5.25}$$

The solution u of the weak formulation can therefore be defined as the element of the space $L^2(\mathbf{curl}, d, \alpha)$ of quasiperiodic functions (i.e. such that $u(x,y) = u_\#(x,y)\, e^{i\alpha x}$ with $u_\#(x,y) = u_\#(x+d,y)$, a d-periodic function) of $L^2(\mathbf{curl})$ on Ω such that:

$$\mathscr{R}_{\underline{\underline{\xi}},\chi}(u,u') = 0 \;\; \forall u' \in L^2(\mathbf{curl}, d, \alpha). \tag{5.26}$$

As for the boundary term introduced by the integration by part, it can be classically set to zero by imposing Dirichlet conditions on a part of the boundary (the value of u is imposed and the weight function u' can be chosen equal to zero on this part of the boundary) or by imposing homogeneous Neumann conditions $(\underline{\underline{\xi}} \nabla u) \cdot \mathbf{n} = 0$ on another part of the boundary (and u is

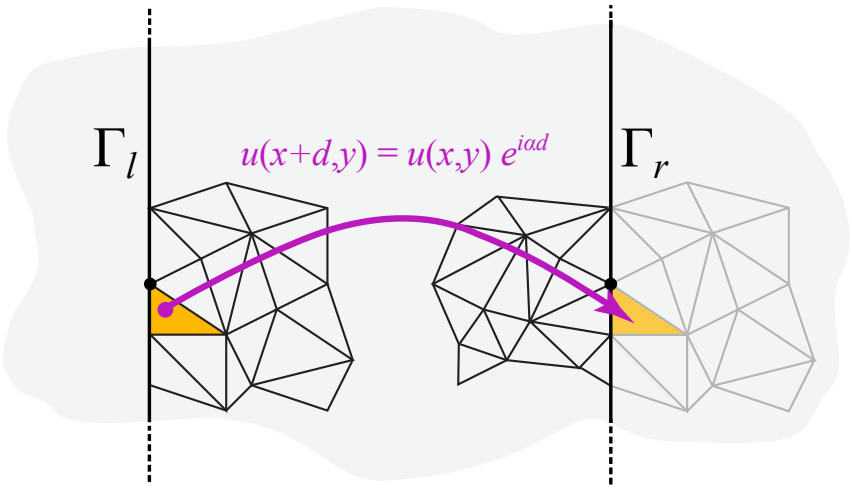


Fig. 5.2: *Quasi-periodicity of the field and sample of a d-periodic mesh.*

therefore an unknown to be determined on the boundary). A third possibility are the so-called quasi-periodicity conditions of particular importance in the modeling of gratings.

Denote by Γ_l and Γ_r the lines parallel to the y–axis delimiting a cell of the grating (see Fig. 5.2) respectively from its left and right neighbor cell. Considering that both u and u' are in $L^2(\mathbf{curl}, d, \alpha)$, the boundary term for $\Gamma_l \cup \Gamma_r$ is

$$\int_{\Gamma_l \cup \Gamma_r} \overline{u'} \left(\underline{\underline{\xi}} \nabla u \right) \cdot \mathbf{n} \, \mathrm{d}S = \int_{\Gamma_l \cup \Gamma_r} \overline{u'_\#} e^{-i\alpha x} \left(\underline{\underline{\xi}} \nabla (u_\# e^{+i\alpha x}) \right) \cdot \mathbf{n} \, \mathrm{d}S =$$

$$\int_{\Gamma_l \cup \Gamma_r} \overline{u'_\#} \left(\underline{\underline{\xi}} \left(\nabla u_\# + i\alpha u_\# \mathbf{x} \right) \right) \cdot \mathbf{n} \, \mathrm{d}S = 0 \,,$$

because the integrand $\overline{u'_\#} \left(\underline{\underline{\xi}} \left(\nabla u_\# + i\alpha u_\# \mathbf{x} \right) \right) \cdot \mathbf{n}$ is periodic along x and the normal $\mathbf{n}$ has opposite directions on Γ_l and Γ_r so that the contributions of these two boundaries have the same absolute value with opposite signs. The contribution of the boundary terms vanishes therefore naturally in the case of quasi-periodicity.

The finite element method is based on this weak formulation and both the solution and the weight functions are classically chosen in a discrete space made of linear or quadratic Lagrange elements, i.e. piecewise first or second order two variable polynomial interpolation built on a triangular mesh of the domain Ω (cf. Fig. 5.3a). Dirichlet and Neumann conditions may be used to truncate the PML domain in a region where the field (transformed by the PML) is negligible. The quasi-periodic boundary conditions are imposed by considering the u as unknown on Γ_l (in a way similar to the homogeneous Neumann condition case) while, on Γ_r, u is forced equal to the value of the corresponding point on Γ_l (i.e. shifted by a quantity $-d$ along x) up to the factor $e^{i\alpha d}$. The practical implementation in the finite element method is described in details in [10, 11]

5.2.2.4 *Perfectly Matched Layer for z–anisotropic materials*

The main drawback encountered in electromagnetism when tackling theory of gratings through the finite element method is the non-decreasing behaviour of the propagating modes in superstratum and substratum (if they are made of lossless materials): The PML has been introduced

by [4] in order to get round this obstacle. The computation of PML designed for z–anisotropic gratings is the topic of what follows.

In the framework of transformation optics, a PML may be seen as a change of coordinate corresponding to a *complex stretch* of the coordinate corresponding to the direction along which the field must decay [12, 13, 14]. Transformation optics have recently unified various techniques in computational electromagnetics such as the treatment of open problems, helicoidal geometries or the design of invisibility cloaks ([15]). These apparently different problems share the same concept of geometrical transformation, leading to equivalent material properties. A very simple and practical rule can be set up ([10]): when changing the coordinate system, all you have to do is to replace the initial materials properties $\underline{\underline{\varepsilon}}$ and $\underline{\underline{\mu}}$ by equivalent material properties $\underline{\underline{\varepsilon}}_s$ and $\underline{\underline{\mu}}_s$ given by the following rule:

$$\underline{\underline{\varepsilon}}_s = \mathbf{J}^{-1}\,\underline{\underline{\varepsilon}}\,\mathbf{J}^{-\mathrm{T}}\det(\mathbf{J}) \;\text{ and }\; \underline{\underline{\mu}}_s = \mathbf{J}^{-1}\,\underline{\underline{\mu}}\,\mathbf{J}^{-\mathrm{T}}\det(\mathbf{J}), \tag{5.27}$$

where $\mathbf{J}$ is the Jacobian matrix of the coordinate transformation consisting of the partial derivatives of the new coordinates with respect to the original ones ($\mathbf{J}^{-\mathrm{T}}$ is the transposed of its inverse).
In this framework, the most natural way to define PMLs is to consider them as maps on a complex space $\mathbb{C}^3$, which coordinate change leads to equivalent permittivity and permeability tensors. We detail here the different coordinates used in this section.

- (x,y,z) are the cartesian original coordinates.
- (x_s,y_s,z_s) are the complex stretched coordinates. A suitable subspace $\Gamma \subset \mathbb{C}^3$ is chosen (with three real dimensions) such that (x_s,y_s,z_s) are the complex valued coordinates of a point on Γ (e.g. $x = \mathfrak{Re}(x_s)$, $y = \mathfrak{Re}(y_s)$, $z = \mathfrak{Re}(z_s)$).
- (x_c,y_c,z_c) are three real coordinates corresponding to a real valued parametrization of $\Gamma \subset \mathbb{C}^3$.

We use rectangular PMLs ([12]) absorbing in the y-direction and we choose a diagonal matrix $\mathbf{J} = \mathrm{diag}(1, s_y(y), 1)$, where $s_y(y)$ is a complex-valued function of the real variable y, defined by:

$$y_s(y) = \int_0^y s_y(y')\mathrm{d}y'. \tag{5.28}$$

The expression of the equivalent permittivity and permeability tensors are thus:

$$\underline{\underline{\varepsilon}}_s = \begin{pmatrix} s_y\varepsilon_{xx} & \overline{\varepsilon_a} & 0 \\ \varepsilon_a & s_y^{-1}\varepsilon_{yy} & 0 \\ 0 & 0 & s_y\varepsilon_{zz} \end{pmatrix} \quad \text{and} \quad \underline{\underline{\mu}}_s = \begin{pmatrix} s_y\mu_{xx} & \overline{\mu_a} & 0 \\ \mu_a & s_y^{-1}\mu_{yy} & 0 \\ 0 & 0 & s_y\mu_{zz} \end{pmatrix}. \tag{5.29}$$

Note that the equivalent medium has the same impedance than the original one as $\underline{\underline{\varepsilon}}$ an $\underline{\underline{\mu}}$ are transformed in the same way, which guarantees that the PML is perfectly reflectionless.
Now, let us define the so-called substituted field $\boldsymbol{F}_s = (\mathbf{E}_s, \mathbf{H}_s)$, solution of Eqs. (5.3) with $\underline{\underline{\xi}} = \underline{\underline{\xi}}_s$ and $\underline{\underline{\chi}} = \underline{\underline{\chi}}_s$. It turns out that $\boldsymbol{F}_s$ equals the field $\boldsymbol{F}$ in the region $y^b < y < y^t$ (with $y^b = -h^-$ and $y^t = h^g + h^+$, see Fig. 5.3a), provided that $s_y(y) = 1$ in this region. The main feature of this latest field $\boldsymbol{F}_s$ is the remarkable correspondence with the first field $\boldsymbol{F}$; whatever the function s_y provided that it equals 1 for $y^t < y < y^b$, the two fields $\boldsymbol{F}$ and $\boldsymbol{F}_s$ are identical in the region $y^t < y < y^b$[8]. In other words, the PML is completely reflection-less. In addition, for complex

valued functions s_y ($\Im m\{s_y\}$ strictly positive in PML), the field $\boldsymbol{F}_s$ converges exponentially towards zero (as y tends to $\pm\infty$, cf. Fig. 5.3c and 5.3d) although its physical counterpart $\boldsymbol{F}$ does not. Note that in Fig. 5.3d, the value of the computed radiated field u_2^d on each extreme boundary of the PMLs is at least 10^{-8} weaker than in the region of interest. As a consequence, $\boldsymbol{F}_s$ is of finite energy and for this substituted field a weak formulation can be easily derived which is essential when dealing with Finite Element Method.

Still remains to give a suitable function s_y. Let us consider the complex coordinate mapping $y(y_c)$, which is simply defined as the derivative of the stretching coefficient $s_y(y)$ with respect to y_c. With simple stretching functions, we can obtain a reliable criterion upon proper fields decay. A classical choice is:

$$s_y(y) = \begin{cases} \zeta^- & \text{if } y < y^b \\ 1 & \text{if } y^b < y < y^t \\ \zeta^+ & \text{if } y > y^t \end{cases} \tag{5.30}$$

where $\zeta^\pm = \zeta^{\prime,\pm} + i\zeta^{\prime\prime,\pm}$ are complex constants with $\zeta^{\prime\prime,\pm} > 0$.
In that case, the complex valued function $y(y_c)$ defined by Eq. (5.28) is explicitly given by:

$$y(y_c) = \begin{cases} y^b + \zeta^-(y_c - y^b) & \text{if } y_c < y^b \\ y_c & \text{if } y^b < y_c < y^t \\ y^t + \zeta^+(y_c - y^t) & \text{if } y_c > y^t \end{cases}, \tag{5.31}$$

Finally, let us consider a propagating plane wave in the substratum $u_n(x,y) := \exp(i(\alpha x - \beta_n^- y))$. Its expression can be rewritten as a function of the stretched coordinates in the PML as follows:

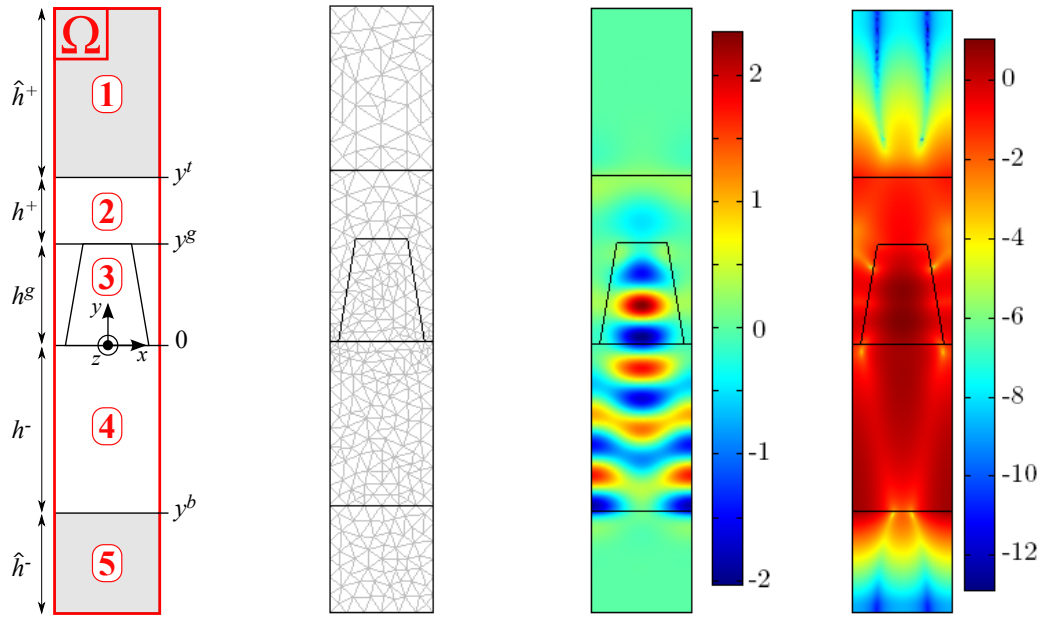
$$u_n^{\text{sc}}(x_c, y_c) := u_n(x(x_c), y(y_c)) = e^{i\alpha x_c} e^{-i\beta_n^-(y^b + \zeta^-(y_c - y^b))} \tag{5.32}$$

The behavior of this latest function along the y_c direction is governed by the function $U^{\text{sc}}(y_c) := e^{-i\beta_n^- \zeta^- y_c}$. Letting $\beta_n^{\prime,-} := \Re e\{\beta_n^-\}$, $\beta_n^{\prime\prime,-} := \Im m\{\beta_n^-\}$, $\zeta^{\prime,-} := \Re e\{\zeta^-\}$ and $\zeta^{\prime\prime,-} := \Im m\{\zeta^-\}$, the non-oscillating part of the function $U^{\text{sc}}(y_c)$ is given by $\exp\Big((\beta_n^{\prime,-}\,\zeta^{\prime\prime,-} + \beta_n^{\prime\prime,-}\,\zeta^{\prime,-})y_c\Big)$. Keeping in mind that $\beta_n^{\prime,-}$ and/or $\beta_n^{\prime\prime,-}$ are positive numbers, the function U^{sc} decreases exponentially towards zero as y_c tends to $-\infty$ (Fig. 5.3d) provided that ζ^- belongs to $\mathbb{C}^+ := \{z \in \mathbb{C}, \Re e\{z\} > 0, \text{ and } \Im m\{z\} > 0\}$. In the same way, it can be shown that ζ^+ belongs to $\mathbb{C}^+$.

Let us conclude this section with two important remarks:

1. **Practical choice of PML parameters.** As for the complex stretch parameters, setting $\zeta^\pm = 1 + i$ is usually a safe choice. For computational needs, the PML has to be truncated and the other constitutive parameter of the PML is its thickness $\hat{h}$ (see Fig. 5.3a). Setting $\hat{h}^\pm = \lambda_0/\sqrt{\varepsilon^\pm}$ leads to a PML thick enough to "absorb" all incident radiation. These specific values will be used in the sequel, unless otherwise specified.

2. **Special cases.** The reader will notice that a configuration where $\beta_n^{\prime,-}$ is a very weak positive number compared to k_0 with $\beta_n^{\prime\prime,-}$ (this is precisely the case of a plane wave at grazing incidence on the bottom PML) **leads to a very slow exponential decay** of U^{sc}. In such a case, close to so-called Wood's anomalies or at extreme grazing incidences, classical PML fail. We will address this tricky situation extensively in Section 5.2.4.



(a) *Computational domain Ω and its five constituent regions.*

(b) *Coarse triangle meshing of the cell Ω. Maximum element side size:* $\lambda/(2\sqrt{\varepsilon})$

(c) *Radiated field:* $\Re e\{u_2^d\}$ *in* V/m

(d) *Radiated field:* $\log(|u_2^d|)$

Fig. 5.3: *Example of computation of the radiated field* u_2^d *(TM case).*

5.2.2.5 Synthesis of the method

In order to give a general view of the method, all information is collected here that is necessary to set up the practical Finite Element Model. First of all, the computation domain Ω (cf. Fig. 5.3a) corresponds to a truncated cell of the grating which is a finite rectangle divided into five horizontal layers. These layers are respectively from top to bottom upper PML, the superstratum, the groove region, the substratum, and the lower PML. The unknown field is the scalar function u_2^d defined in Eq. (5.16). Its finite element approximation is based on the second Lagrange elements built on a triangle meshing of the cell (cf. Fig. 5.3b). A complex algebraic system of linear equations is constructed via the Galerkin weighted residual method, *i.e.* the set of weight functions u' is chosen as the set of shape functions of interpolation on the mesh [10].

- In region 1 (upper PML, see Fig. 5.3a),

$$\mathscr{R}_{\underline{\underline{\xi}}_s^+,\chi_s^+}(u_2^d,u') = 0\,, \tag{5.33}$$

  with $\underline{\underline{\xi}}_s^+$ and χ_s^+ depending on the equivalent anisotropic properties of the PML given by Eq. (5.7), Eq. (5.8) and Eqs. (5.29).

- In region 2 (superstratum),

$$\mathscr{R}_{\underline{\underline{\xi}}^+,\chi^+}(u_2^d,u') = 0\,, \tag{5.34}$$

  with $\underline{\underline{\xi}}^+$ and χ^+ depending on the homogeneous isotropic properties of the superstratum given by Eq. (5.7), Eq. (5.8), Eq. (5.9) and Eq. (5.10).

- In region 3 (groove region),

$$\mathscr{R}_{\underline{\underline{\xi}}^g,\chi^g}(u_2^d,u') = -\mathscr{R}_{\underline{\underline{\xi}}^g,\chi^g}(\mathscr{S}_1,u') \,, \tag{5.35}$$

with $\underline{\underline{\xi}}^g$ and χ^g depending on the heterogeneous possibly anisotropic properties given by Eq. (5.7), Eq. (5.8), Eq. (5.11) and $\mathscr{S}_1$ given by Eq. (5.19) , Eq. (5.22), Eq. (5.23) and Eq. (5.24).

- In region 4 (substratum),

$$\mathscr{R}_{\underline{\underline{\xi}}^-,\chi^-}(u_2^d,u') = 0 \,, \tag{5.36}$$

with $\underline{\underline{\xi}}^-$ and χ^- depending on the homogeneous isotropic properties of the substratum given by Eq. (5.7), Eq. (5.8), Eq. (5.9) and Eq. (5.10).

- In region 5 (lower PML),

$$\mathscr{R}_{\underline{\underline{\xi}}^-_s,\chi_s^-}(u_2^d,u') = 0 \,, \tag{5.37}$$

with $\underline{\underline{\xi}}^-_s$ and χ_s^- depending on the equivalent anisotropic properties of the PML given by Eq. (5.7), Eq. (5.8) and Eqs. (5.29).

5.2.2.6 Energy balance: Diffraction efficiencies and absorption

The rough result of the FEM calculation is the complex *radiated* field u_2^d. Using Eq. (5.16), it is straightforward to obtain the complex *diffracted* field u^d solution of Eq. (5.14) at each point of the bounded domain. We deduce from u^d the diffraction efficiencies with the following method. The superscripts $^+$ (resp. $^-$) correspond to quantities defined in the superstratum (resp. substratum) as previously.

On the one hand, since u^d is quasi-periodic along the x–axis , it can be expanded as a Rayleigh expansion (see for instance [9]):

$$\text{for } y<0 \text{ and } y>h^g,\ u^d(x,y) = \sum_{n\in\mathbb{Z}} u_n^d(y)\, e^{i\alpha_n x} \tag{5.38}$$

where

$$u_n^d(y) = \frac{1}{d}\int_{-d/2}^{d/2} u^d(x,y) e^{-i\alpha_n x}\,\mathrm{d}x \text{ with } \alpha_n = \alpha + \frac{2\pi}{d}n \tag{5.39}$$

On the other hand, introducing Eq. (5.38) into Eq. (5.6) leads to the Rayleigh coefficients:

$$u_n^d(y) = \begin{cases} u_n^+(y) = r_n\, e^{i\beta_n^+ y} + a_n\, e^{-i\beta_n^+ y} & \text{for} \quad y > h^g \\ \\ u_n^-(y) = t_n\, e^{-i\beta_n^- y} + b_n\, e^{i\beta_n^- y} & \text{for} \quad y < 0 \end{cases} \quad \text{with } \beta_n^{\pm^2} = k^{\pm^2} - \alpha_n^2 \tag{5.40}$$

For a temporal dependance in $e^{-i\omega t}$, the O.W.C. imposes $a_n = b_n = 0$. Combining Eq. (5.39) and (5.40) at a fixed y_0 altitude leads to:

$$\begin{cases} r_n = \frac{1}{d}\displaystyle\int_{-d/2}^{d/2} u^d(x,y_0)\, e^{-i(\alpha_n x + \beta_n^+ y_0)}\,\mathrm{d}x \quad \text{for} \quad y_0 > h^g \\ \\ t_n = \frac{1}{d}\displaystyle\int_{-d/2}^{d/2} u^d(x,y_0)\, e^{-i(\alpha_n x - \beta_n^- y_0)}\,\mathrm{d}x \quad \text{for} \quad y_0 < 0 \end{cases} \tag{5.41}$$

We extract these two coefficients by trapezoidal numerical integration along x from a cutting of the previously calculated field map at y_0. It is well known that the mere trapezoidal integration method is very efficient for smooth and periodic functions (integration on one period) [16]. Now the restriction on a horizontal straight line crossing the whole cell in homogeneous media (substratum and superstratum) is of C^∞ class. From a numerical point of view, it appears that the interpolated approximation of the unknown function, namely u_2^d preserves the good behaviour of the numerical computation of these integrals. From this we immediately deduce the reflected and transmitted diffracted efficiencies of propagative orders (T_n and R_n) defined by:

$$\left\{ \begin{array}{ll} R_n := r_n \overline{r_n} \frac{\beta_n^+}{\beta^+} & \text{for} \quad y_0 > h^g \\ \\ T_n := t_n \overline{t_n} \frac{\beta_n^-}{\beta^-} \frac{\gamma^+}{\gamma^-} & \text{for} \quad y_0 < 0 \end{array} \right. \quad \text{with } \gamma^\pm = \left\{ \begin{array}{l} 1 \text{ in the TM case} \\ \\ \varepsilon^\pm \text{ in the TE case} \end{array} \right. \tag{5.42}$$

This calculation is performed at several different y_0 altitudes in the superstratum and the substratum, and the mean value found for each propagative transmitted or reflected diffraction order is presented in the numerical experiments of the following section.

Normalized losses Q can be obtained according to Poynting's theorem through the straightforward computation of the following ratio:

$$Q := \frac{\displaystyle\int_S \omega \varepsilon_0 \Im m(\varepsilon^{g'}) \mathbf{E} \cdot \overline{\mathbf{E}} \, \mathrm{d}s}{\displaystyle\int_L \Re e\{\mathbf{E}_0 \times \overline{\mathbf{H}_0}\} \cdot \mathbf{n} \, \mathrm{d}l}, \tag{5.43}$$

The numerator in Eq. (5.43) clarifies losses in Watts by period of the considered grating and are computed by integrating the Joule effect losses density over the surface S of the lossy element. The denominator normalizes these losses to the incident power, *i.e.* the time-averaged incident Poynting vector flux across one period (a straight line L of length d in the superstrate parallel to Ox, whose normal oriented along decreasing values of y is denoted $\mathbf{n}$).

Finally, combining Eqs. (5.42) and Eq. (5.43), a self consistency check of the whole numerical scheme consists in comparing the quantity B:

$$B := \sum_n T_n + \sum_m R_m + Q \tag{5.44}$$

to unity. In Eq. (5.44), the summation indexed by n (resp. m) corresponds to the sum over the efficiencies of all transmitted (resp. reflected) propagative diffraction orders in the substrate (resp. superstrate). We give interpretations and concrete examples of such numerical energy balances over non trivial grating profiles in sections 5.2.3.2 and 5.2.3.3.

5.2.3 Numerical experiments

5.2.3.1 Numerical validation of the method

We can refer to [17] in order to test the accuracy of our method. The studied grating is isotropic, since we lack numerical values in the literature in anisotropic cases. We compute the following problem (cf. Fig. 5.4), as described in [18] and [17]. The wavelength of the plane wave is set to $1\,\mu m$ and is incoming with an angle of $\pi/6$ with respect to the normal to the grating.

We present the R_0 efficiency (cf. Table 5.1) in both cases of polarization versus the mesh refinement. So we have a good agreement to the reference values, and the accuracy reached is independent from the polarization case.

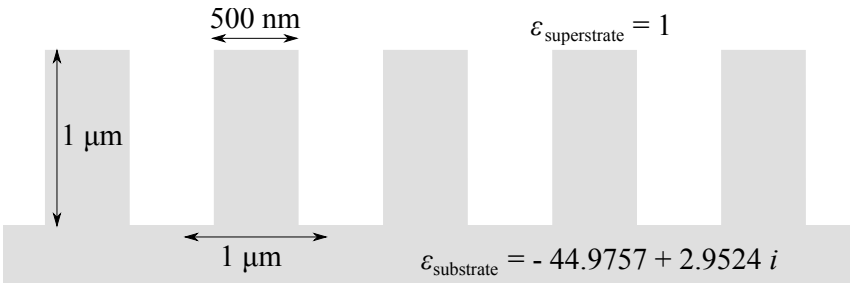


Fig. 5.4: *Rectangular groove grating: This pattern is repeatedly set up with a period d = 1 µm. This grating has been studied by [17] and is one of our points of reference*

Maximum element size	R_0^{TE}	R_0^{TM}
$\lambda_0/(4\sqrt{\varepsilon})$	0.7336765	0.8532342
$\lambda_0/(6\sqrt{\varepsilon})$	0.7371302	0.8456592
$\lambda_0/(8\sqrt{\varepsilon})$	0.7347466	0.8482817
$\lambda_0/(10\sqrt{\varepsilon})$	0.7333739	0.850071
$\lambda_0/(12\sqrt{\varepsilon})$	0.7346569	0.8494844
$\lambda_0/(14\sqrt{\varepsilon})$	0.7341944	0.8483238
$\lambda_0/(16\sqrt{\varepsilon})$	0.7342714	0.8484774
Result given by [17]	0.7342789	0.8484781

Tab. 5.1: *Reflected efficiencies versus mesh refinement. Note that the efficiencies are properly computed (two significant digits) even for a rather coarse mesh.*

5.2.3.2 Experiment set up based on existing materials

The method proposed in this section is adapted to z–anisotropic materials, such as transparent $CaCO_3$ [19], $LiNbO_3$ [20] or Ni:YIG [21] and lossy CoPt or CoPd [22]. Let us now consider a trapezoidal (cf. Fig. 5.5) anisotropic grating made of aragonite ($CaCO_3$) deposited on an isotropic substratum (SiO_2, $\varepsilon_{SiO_2} = 2.25$). Along the anisotropic crystal axis, its dielectric tensor can be written as follows [19]:

$$\underline{\underline{\varepsilon}}_{\text{CaCO}_3} = \begin{pmatrix} 2.843 & 0 & 0 \\ 0 & 2.341 & 0 \\ 0 & 0 & 2.829 \end{pmatrix} \quad \text{and} \quad \underline{\underline{\mu}}_{\text{CaCO}_3} = \begin{pmatrix} \mu_0 & 0 & 0 \\ 0 & \mu_0 & 0 \\ 0 & 0 & \mu_0 \end{pmatrix} \tag{5.45}$$

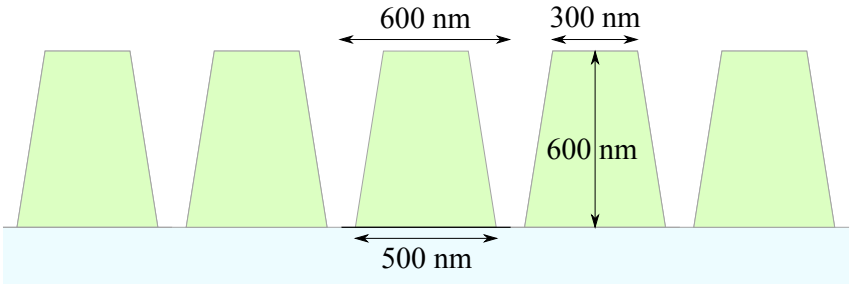


Fig. 5.5: *Diffractive element pattern. This element is made of aragonite for which the dielectric tensor is given by Eq. (5.46) and is deposited on a silica substrate with a period $d = 600nm$.*

Now let's assume that the natural axis of our aragonite grating are rotated by 45° around

the grating infinite dimension. The dielectric tensor becomes:

$$\underline{\underline{\varepsilon}}^{45^\circ}_{CaCO_3} = \begin{pmatrix} 2.592 & 0.251 & 0 \\ 0.251 & 2.592 & 0 \\ 0 & 0 & 2.829 \end{pmatrix} \tag{5.46}$$

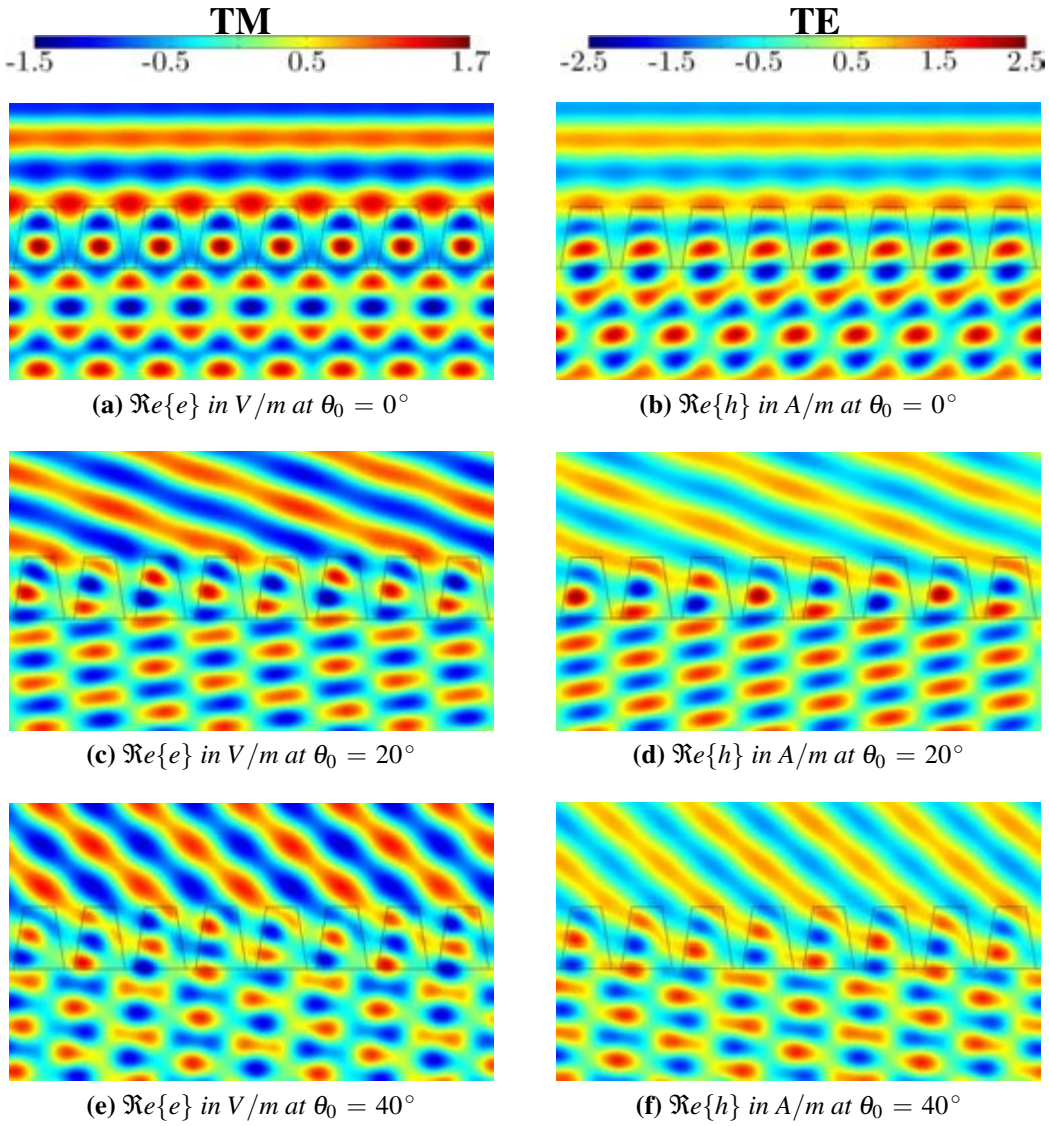
We shall here remind that our method remains strictly the same whatever the diffractive element geometry is. The 2D computational domain is bounded along the y–axis by the PMLs and along the x since we consider only one pseudo period. We propose to calculate the diffractive efficiencies at $\lambda_0 = 633\,\text{nm}$ in both polarization cases TE and TM, and for different incoming incidences (0°, 20° and 40°). Since both $\underline{\underline{\mu}}$ and $\underline{\underline{\varepsilon}}$ are Hermitian, the whole incident energy is diffracted and the sum of theses efficiencies ought to be equal to the incident energy, which will stand for validation of our numerical calculation.

Finally, the resulting bounded domain is meshed with a maximum mesh element side size of $\lambda_0/10\sqrt{\varepsilon}$. Efficiencies are still post-processed in accordance with the calculation presented section 5.2.2.6.

TM	T_{-2}	T_{-1}	T_0	T_1	R_{-1}	R_0	R_1	total
0°	-	0.203133	0.585235	0.203138	-	0.008473	-	0.999978
20°	-	0.399719	0.575625	0.004643	0.004412	0.015630	-	1.000029
40°	0.025047	0.420714	0.493491	-	0.002541	0.058238	-	1.000031
TE	T_{-2}	T_{-1}	T_0	T_1	R_{-1}	R_0	R_1	total
0°	-	0.322510	0.538165	0.124722	-	0.014683	-	1.000080
20°	-	0.538727	0.444403	0.000369	0.005372	0.011180	-	1.000051
40°	0.012058	0.434191	0.541090	-	0.005032	0.007686	-	1.000057

Tab. 5.2: *Transmitted and reflected efficiencies of propagative orders deduced from field maps shown Fig. 5.6*

At normal incidence, the h field in the TE case (cf. Fig. 5.6b) is non symmetric whereas the e field in the TM case is (cf. Fig. 5.6a). This is illustrated by the obvious non-symmetry of T_{-1}^{TE} and T_1^{TE} (cf. Table 5.2: 0.322510 versus 0.124722!), whereas $T_{-1}^{\text{TM}} = T_1^{\text{TM}} = 0.20313$.



(a) $\Re e\{e\}$ *in* V/m *at* $\theta_0 = 0°$ **(b)** $\Re e\{h\}$ *in* A/m *at* $\theta_0 = 0°$

(c) $\Re e\{e\}$ *in* V/m *at* $\theta_0 = 20°$ **(d)** $\Re e\{h\}$ *in* A/m *at* $\theta_0 = 20°$

(e) $\Re e\{e\}$ *in* V/m *at* $\theta_0 = 40°$ **(f)** $\Re e\{h\}$ *in* A/m *at* $\theta_0 = 40°$

Fig. 5.6: *Real part of the total calculated field depending on* θ_0 *and the polarization case*

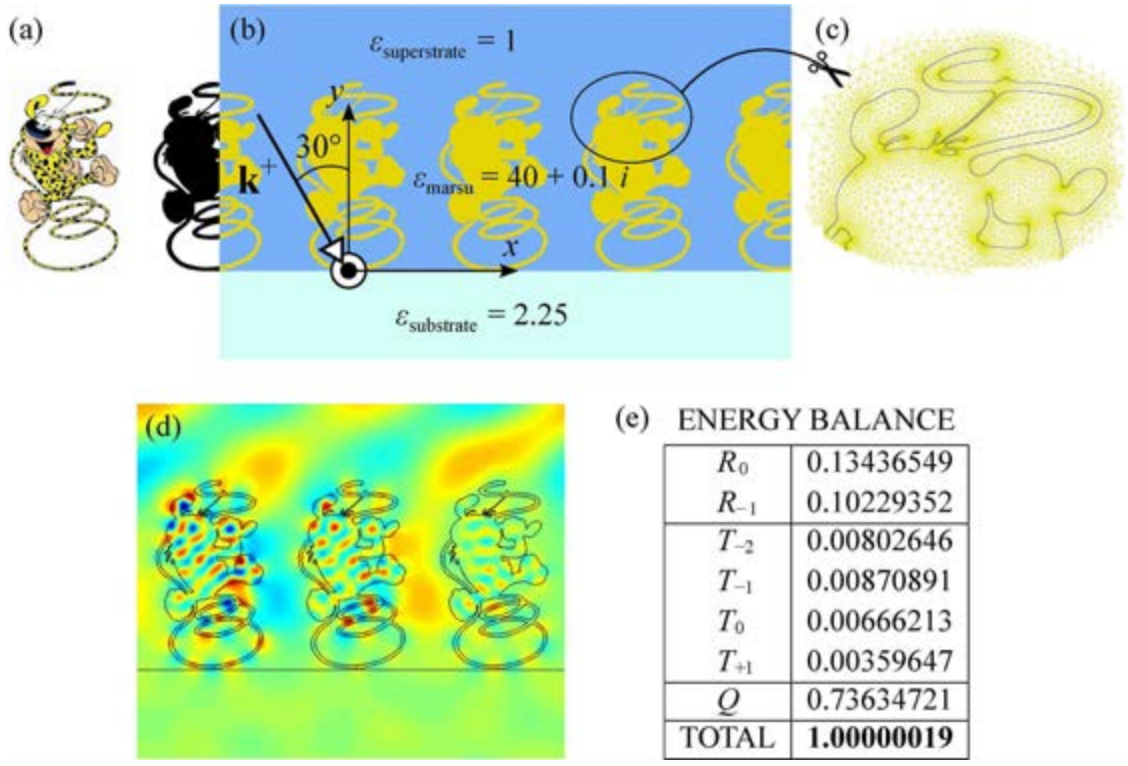
5.2.3.3 A non trivial geometry

Since the beginning of this chapter, we have laid great stress upon the independence of the method towards the geometry of the pattern. But we have considered so far diffractive objects of simple trapezoidal section. Let us tackle a way more challenging case and see what this approach is made of.

We can obtain an quite winding shape by extracting the contrast contour of an arbitrary image (see Fig. 5.7a-5.7b). The contour is approximated by a set of splines, and the resulting domain is finely meshed (Fig. 5.7c). Finally, as shown in Fig. 5.7b, the formed pattern ($h^g/\lambda_0 = 1.68$), breathing in free space ($\varepsilon_{\text{substrat}} = 1$), is supposed to be periodically repeated $d/\lambda_0 = 1.26$ on a plane ground of glass ($\varepsilon_{SiO_2} = 2.25$). The element is considered to be "made of" a lossy material of high optical index ($\varepsilon_{\text{marsu}} = 40 + 0.1\,i$). The response of this system to a incident s-polarized plane wave at oblic incidence ($\theta_0 = -30°$) is finally calculated. The real part of the quasi-periodic total field is represented in Fig. 5.7d for three periods.

Indeed, we do not have any tabulated data available to check our results. But what we

do have is a pretty reliable consistency check through the computation of the energy balance described by Eqs. (5.42) and (5.43). As shown in Fig. 5.7e, we obtain at least 7 significative digits on the energetic values. The total balance of 1.00000019 is computed taking into account (i) values of the total field inside the diffractive elements, (ii) values of the diffracted field at altitudes spanning the whole (modeled) superstrate, (iii) values of the total field at altitudes spanning the entire (modeled) substrate. Finally, (iv) the calculated field u_2^d also nicely decays exponentially inside both PML. These four points allow us to check *a posteriori* the validity of the field everywhere in the computation cell.



(e) ENERGY BALANCE

R_0	0.13436549
R_{-1}	0.10229352
T_{-2}	0.00802646
T_{-1}	0.00870891
T_0	0.00666213
T_{+1}	0.00359647
Q	0.73634721
TOTAL	**1.00000019**

Fig. 5.7: *(a) Initial contrasted image. (b) Proposed set up. (c) Sample mesh. (d)* $\Re e\{E_z\}$ *in* V/m*. (e) Energy balance of the problem.*

5.2.4 Dealing with Wood anomalies using Adaptative PML

As we have noticed at the end of Section 5.2.2.4, PMLs based on "traditional coordinate stretching" are inefficient for periodic problems when dealing with grazing angles of diffracted orders, *i.e.* when the frequency is near a Wood's anomaly ([23, 24]), leading to spurious reflexions and thus numerical pollution of the results. An important question in designing absorbing layers is thus the choice of their parameters: The PML thickness and the absorption coefficient. To this aim, adaptative formulations have already been set up, most of them employing a posteriori error estimate [25, 18, 26]. In this section, we propose Adaptative PMLs (APMLs) with a suitable coordinate stretching, depending both on incidence and grating parameters, capable of efficiently absorbing propagating waves with nearly grazing angles. This section is dedicated to the mathematical formulation used to determine PML parameters adapted to any diffraction orders. We provide at the end a numerical example of a dielectric slit grating showing the relevance of our approach in comparison with classical PMLs.

5.2.4.1 Skin depth of the PML

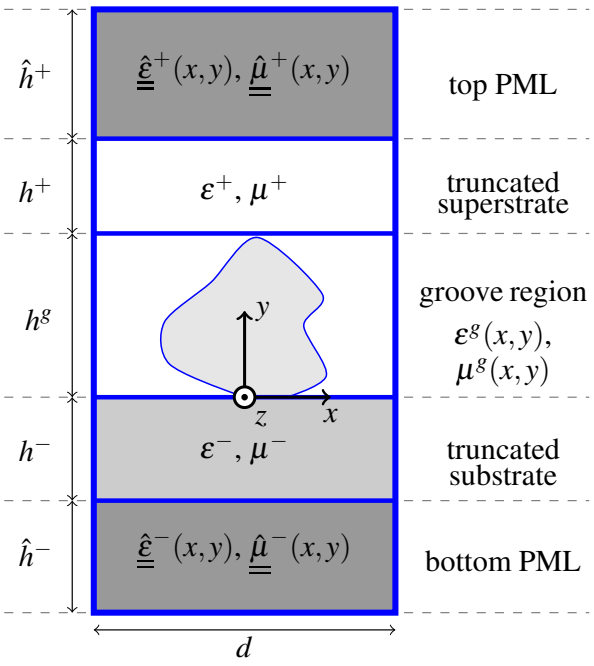


Fig. 5.8: *The basic cell used for the FEM computation of the diffracted field* u_2^d.

As explained in Section 5.2.2.6, the diffracted field u^d can be expanded as a Rayleigh expansion, *i.e.* into an infinite sum of propagating and evanescent plane waves called diffraction orders. As detailed at the end of Section 5.2.2.4, we are now in position to rewrite easily the expression of, say, a transmitted diffraction order into the substrate. Similar considerations also apply to the reflected orders in the top PML. Combining Eq. (5.32) and (5.40) lead to the expression $u_{n,s}^-(y_c)$ of a transmitted propagative order inside the PML:

$$u_{n,s}^-(y_c) = u_n^-(y(y_c)) = t_n \mathbf{e}^{-i\beta_n^-[y^t+\zeta^-(y_c-y^t)]}.$$

The non oscillating part of this function is given by:

$$U_n^-(y) = t_n \exp\left((\beta_n'^{,-}\zeta''^{,-} + \beta_n''^{,-}\zeta'^{,-})y_c\right),$$

where $\beta_n^- = \beta_n'^{,-} + i\beta_n''^{,-}$. For a propagating order we have $\beta_n'^{,-} > 0$ and $\beta_n'^{,-} = 0$, while for an evanescent order $\beta_n'^{,-} = 0$ and $\beta_n''^{,-} > 0$. It is thus sufficient to take $\zeta'^{,-} > 0$ and $\zeta''^{,-} > 0$ to ensure the exponential decay to zero of the field inside the PML *if it was of infinite extent*. But, of course, for practical purposes, the thickness of the PML is finite and has to be suitably chosen. Two pitfalls must be avoided:

1. The PML thickness is chosen too small compared to the skin depth. As a consequence, the electromagnetic wave cannot be considered as vanishing: An incident electromagnetic "sees the bottom of the PML". In other words, this PML of finite thickness is no longer reflection-less.

2. The PML thickness is chosen much larger than the skin depth. In that case, a significant part of the PML is not useful, which gives rise to the resolution of linear systems of unnecessarily large dimensions.

Then remains to derive the skin depth, l_n^-, associated with the propagating order n. This characteristic length is defined as the depth below the PML at which the field falls to $1/e$ of its value near the surface:

$$U_n^-(y - l_n^-) = \frac{U_n^-(y)}{e}.$$

Finally, we find $l_n^- = (\beta_n'^{,-}\zeta''^{,-} + \beta_n''^{,-}\zeta'^{,-})^{-1}$ and we define l^- as the largest value among the l_n^-:

$$l^- = \max_{n\in\mathbb{Z}} l_n^-.$$

The height of the bottom PML region is set to $\hat{h}^- = 10l^-$.

5.2.4.2 Weakness of the classical PML for grazing diffracted angles

Let us consider the (bottom) PML adapted to the substrate. Similar conclusions will hold for the top PML. The efficiency of the classical PML fails for grazing diffracted angles, in other words when a given order appears/vanishes: this is the so-called Wood's anomaly, well known in the grating theory. In mathematical terms, there exists n_0 such that $\beta_{n_0}^- \simeq 0$. The skin depth of the PML then becomes very large. To compensate this, it is tempting to increase the value of $\zeta''^{,-}$, but it would lead to spurious numerical reflections due to an overdamping. For a fixed value of $\hat{h}^-$, if $\zeta''^{,-}$ is too weak, the absorption in the PMLs is insufficient and the wave is reflected on the outward boundary of the PML. To illustrate these typical behaviors (cf. Fig. 5.9), we compute the field diffracted by a grating with a rectangular cross section of height $h^g = 1.5\,\mu\text{m}$ and width $L^g = 3\,\mu\text{m}$ with $\varepsilon^g = 11.7$, deposited on a substrate with permittivity $\varepsilon^- = 2.25$. The structure is illuminated by a p-polarized plane wave of wavelength $\lambda_0 = 10\,\mu\text{m}$ and of angle of incidence $\theta_0 = 10^\circ$ in the air ($\varepsilon^+ = 1$). All materials are non magnetic ($\mu_r = 1$) and the periodicity of the grating is $d = 4\,\mu\text{m}$. We set $\hat{h}^- = 10l_0^-$ and $\zeta'^{,-} = 1$.

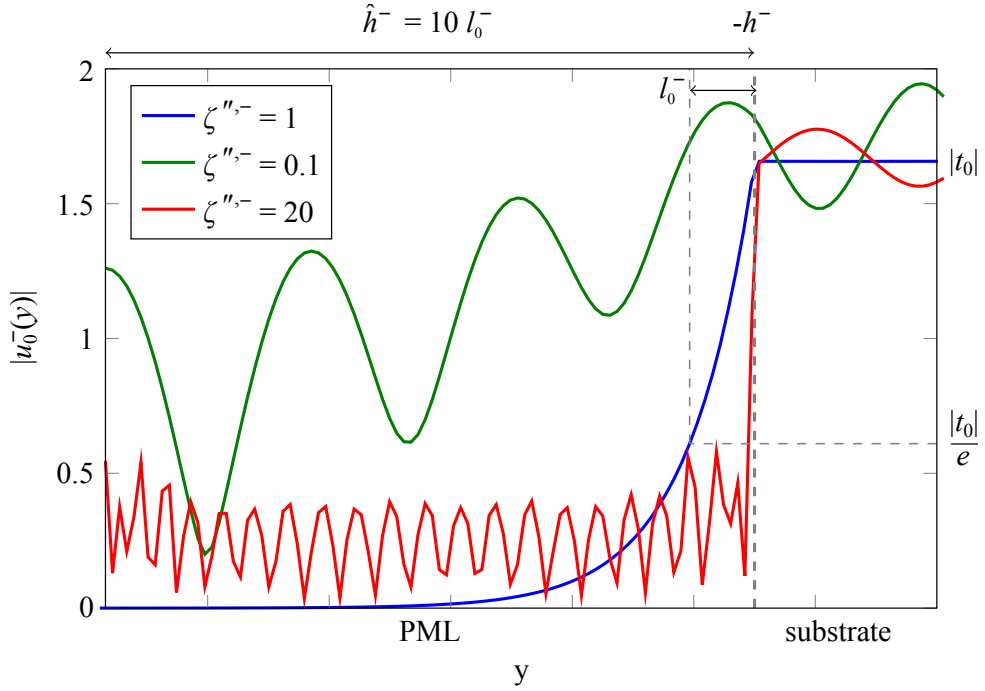


Fig. 5.9: *Zeroth transmitted order by a grating with a rectangular cross section (see parameters in text, part 5.2.4.2) for different values of $\zeta''^{,-}$: blue line, $\zeta''^{,-} = 1$, correct damping; green line, $\zeta''^{,-} = 0.1$, underdamping; red line, $\zeta''^{,-} = 20$, overdamping.*

5.2.4.3 Construction of an adaptative PML

To overcome the problems pointed out in the previous section, we propose a coordinate stretching that rigorously treats the problem of Wood's anomalies. The wavelengths "seen" by the system are very different depending on the order at stake:

- if the diffracted angle θ_n is zero, the apparent wavelength $\lambda_0/\cos\theta_n$ is simply the incident wavelength,
- if the diffracted angle is near $\pm\pi/2$ (grazing angle), the apparent wavelength $\lambda_0/\cos\theta_n$ is very large.

Thus if a classical PML is adapted to one diffracted order, it will not be for another, and vice versa. The idea behind the APML is to deal with each and every order when progressing in the absorbing medium.

Once again the development will be conducted only for the PML adapted to the substrate. We consider a real-valued coordinate mapping $y_d(y)$, the final complex-valued mapping is then $y_c(y) = \zeta^- y_d(y)$, with the complex constant ζ^-, with $\zeta'^{,-} > 0$ and $\zeta''^{,-} > 0$, accounting for the damping of the PML medium.

We begin with transforming the equation $\beta_n^{\pm 2} = k^{\pm 2} - \alpha_n^{\pm 2}$, so that the function with integer argument $n \mapsto \beta_n^-$ becomes a function with real argument continuously interpolated between the imposed integer values. Indeed, the geometric transformations associated to the PML has to be continuous and differentiable in order to compute its Jacobian. To that extent, we choose the parametrization:

$$\alpha(y_d) = \alpha_0 + \frac{2\pi}{d}\frac{y_d}{\lambda_0}, \tag{5.47}$$

so that the application β^- defined by $\beta^-(y_d)^2 = k_0^2\varepsilon^- - \alpha(y_d)^2$ is continuous. Thus, the propagation constant of the n^{th} transmitted order is given by $\beta_n^- = \beta^-(n\lambda_0)$. The key idea is to combine the complex stretching with a real non uniform contraction (given by the continuous function $y(y_d)$, Eq. (5.49)). This contraction is chosen in such a way that for each order n there is a depth y_d^n such that, around this depth, the apparent wavelength corresponding to the order in play is contracted to a value close to λ_0. At that point of the PML, this order is perfectly absorbed thanks to the complex stretch. We thus eliminate first the orders with quasi normal diffracted angles at lowest depths up to grazing orders (near Wood's anomalies) which are absorbed at greater depths. In mathematical words, the translation of previous considerations on the real contraction can be expressed as:

$$\exp\left[-i\beta^-(y_d)y(y_d)\right] = \exp\left(-ik_0 y_d\right) \tag{5.48}$$

The contraction $y(y_d)$ is thus given by:

$$y(y_d) = \frac{k_0 y_d}{\beta^-(y_d)} = \frac{y_d}{\sqrt{\varepsilon^- - (\sin\theta_0 + y_d/d)^2}} \tag{5.49}$$

The function $y(y_d)$ has two poles, denoted $y_{d,\pm}^\star = d(\pm\sqrt{\varepsilon^-} - \sin\theta_0)$. When $y_{d,\pm}^\star = \pm n\lambda_0$ with $n \in \mathbb{N}^\star$, $\beta^-(y_{d,\pm}^\star) = \beta^-(\pm n\lambda_0) = \beta_\pm^- = 0$, i.e. we are on a Wood's anomaly associated with the appearance/disappearance of the $\pm n^{\text{th}}$ transmitted order. We now search for the nearest point to $y_{d,\pm}^*$ associated with a Wood's anomaly, denoting:

$$\begin{cases} n_+^\star / \quad D_+ = \min\limits_{n_+^\star \in \mathbb{N}^\star} |y_{d,+}^\star - n_+^\star \lambda_0| \\ \\ n_-^\star / \quad D_- = \min\limits_{n_-^\star \in \mathbb{N}^\star} |y_{d,-}^\star + n_-^\star \lambda_0| \end{cases} .$$

In a second step, we look for the point $y_d^0 = n^\star\lambda_0$ such that:

$$n^\star / \quad D = \min_{n^\star \in \{n_+^\star, n_-^\star\}} (D^+, D^-) \, . \tag{5.50}$$

To avoid the singular behaviour at $y_d = y_{d,\pm}^\star$, we continue the graph of the function $y_d(y)$ by a straight line tangent at y_d^0, which equation is $t_0(y_d) = s(y_d^0)(y_d - y_d^0) + y(y_d^0)$, where $s(y_d) = \frac{\partial y}{\partial y_d}(y_d)$ is the so-called stretching coefficient. The final change of coordinate is then given by :

$$\tilde{y}(y_d) = \begin{cases} y(y_d) \text{ for } y_d \leq y_d^0 \\ \\ t_0(y_d) \text{ for } y_d > y_d^0. \end{cases} \tag{5.51}$$

Figure 5.10 shows an example of this coordinate mapping. Eventually, the complex stretch s_y used in Eq. (5.29) is given by:

$$s_y(y_d) = \zeta^- \frac{\partial \tilde{y}}{\partial y_d}(y_d). \tag{5.52}$$

Equipped with this mathematical formulation, we can tailor a layer that is doubly perfectly matched:

- to a given medium, which is the aim of the PML technique, through Eq. (5.27),
- to all diffraction orders, through the stretching coefficient s_y, which depends on the characteristics of the incident wave and on opto-geometric parameters of the grating.

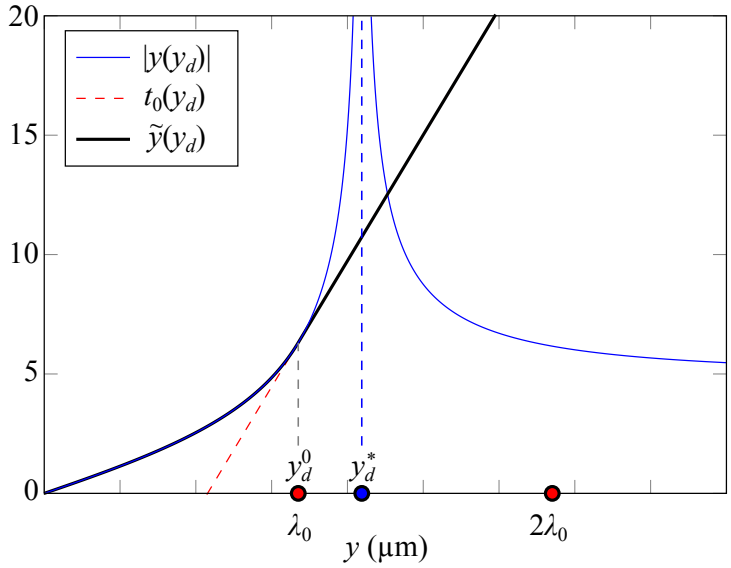


Fig. 5.10: *Example of a coordinate mapping $\tilde{y}(y_d)$ used for the APML (black solid line). The graph of $y_d(y)$ (blue solid line) is continued by a straight line $t_0(y_d)$ tangent at y_d^0 (red dashed line) to avoid the singular behaviour at $y_d = y_d^\star$.*

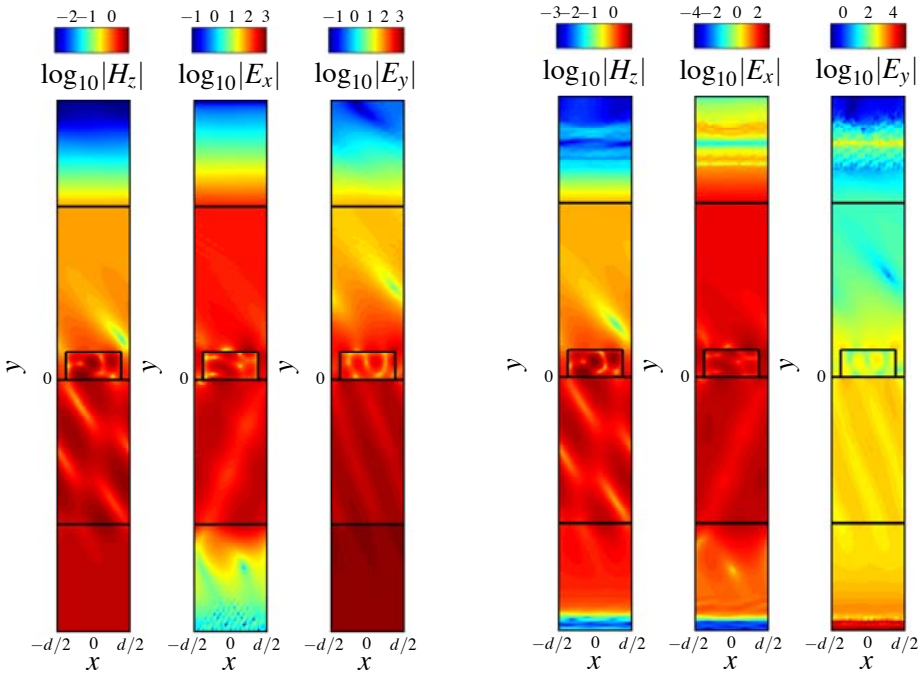
5.2.4.4 Numerical example

We now apply the method described in the preceding parts to design an adapted bottom PML for the same example as in part 5.2.4.2. The parameters are the same, and we choose the wavelength of the incident plane wave close to the Wood's anomaly related to the $+1$ transmitted order ($\lambda_0 = 0.999 y^\star_{d,+}$).Moreover, we set the length of the PML $\hat{h}^- = 1.1 y^\star_{d,+}$ and choose absorption coefficients $\zeta^+ = \zeta^- = 1 + i$. For both cases (PML and APML), parameters are alike, the only difference being the complex stretch s_y.

The field maps of the norm of H_z, E_x and E_y are plotted in logarithmic scale on Fig. 5.11, for the case of a classical PML and our APML. We can observe that the field H_z that is effectively computed is clearly damped in the bottom APML (leftmost on Fig. 5.11(b)) whereas it is not in the standard case (leftmost on Fig. 5.11(a)), causing spurious reflections on the outer boundary. The fields E_x and E_y are deduced from H_z thanks to Maxwell's equations. The high values of E_y at the tip of the APML (rightmost on Fig. 5.11(b)) are due to very high values of the optical equivalent properties of the APML medium (due to high values of s_y), which does not affect the accuracy of the computed field within the domain of interest.

Another feature of our approach is that it efficiently absorbs the grazing diffraction order, as illustrated on Fig. 5.12: the $+1$ transmitted order does not decrease in the standard PML (blue solid line), and reaches a high value at $y = -\hat{h}^-$, whereas the same order tends to zero as $y \to -\hat{h}^-$ in the case of the adapted PML (blue dashed line).

To further validate the accuracy of the method, we compare the diffraction efficiencies computed by our FEM formulation with PML and APML to those obtained by another method. We choose the Rigorous Coupled Wave Analysis (RCWA), also known as the Fourier Modal Method (FMM, [27]). For the chosen parameters, only the 0^{th} order is propagative in reflexion and the orders -1, 0 and $+1$ are non evanescent in transmission. We can also check the energy balance $B = R_0 + T_{-1} + T_0 + T_{+1}$ since there is no lossy medium in our example. Results are reported in Table 5.3, and show a good agreement of the FEM with APML with the results from RCWA. On the contrary, if classical PML are used, the diffraction efficiencies are less accurate



(a) Classical PML (b) Adapted PML

Fig. 5.11: *Field maps of the logarithm of the norm of H_z, E_x and E_y for the dielectric slit grating at $\lambda_0 = 0.999 y^\star_{d,+}$ (same parameters as in part 5.2.4.2). (a): classical PML with inefficient damping of H_z in the bottom PML. (b): APML where the H_z field is correctly damped in the bottom PML. For both cases the thickness of the PML is $\hat{h}^- = 1.1 y^\star_{d,+}$.*

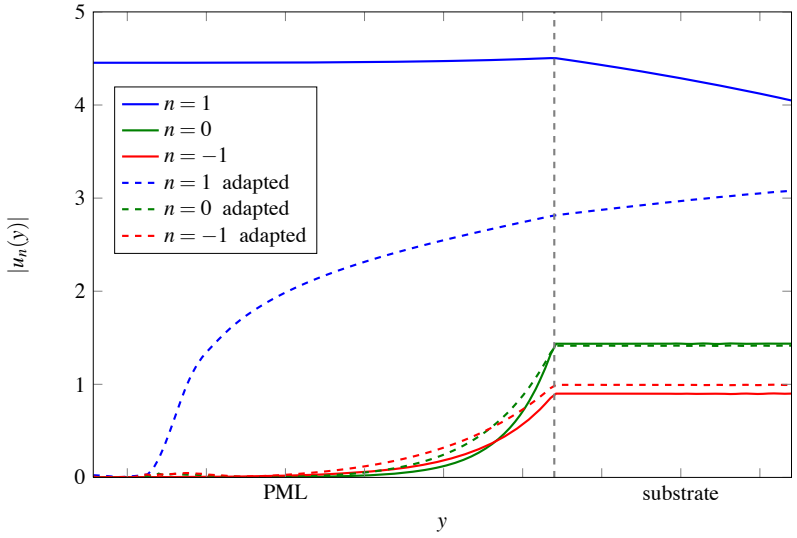


Fig. 5.12: *Modulus of the u_n for the three propagating orders with adapted (dashed lines) and classical PMLs (solid lines). Note that the classical PMLs are efficient for all orders except for the grazing one ($n = 1$) as expected. This drawback is bypassed when using the adaptative PML.*

compared to those computed with RCWA. Checking the energy balance leads the same conclusions: the numerical result is perturbed by the reflection of the waves at the end of the PML if it is not adapted to the situation of nearly grazing diffracted orders.

	R_0	T_{-1}	T_0	T_{+1}	B
RCWA	0.1570	0.3966	0.1783	0.2680	0.9999
FEM + APML	0.1561	0.3959	0.1776	0.2703	0.9999
FEM + PML	0.1904	0.4118	0.1927	0.2481	1.0430

Tab. 5.3: *Diffraction efficiencies R_0, T_{-1}, T_0 and T_{+1} of the four propagating orders, and energy balance $B = R_0 + T_{-1} + T_0 + T_{+1}$, computed by three methods: RCWA (line 1), FEM formulation with APML (line 2), FEM formulation with classical PML (line 3).*

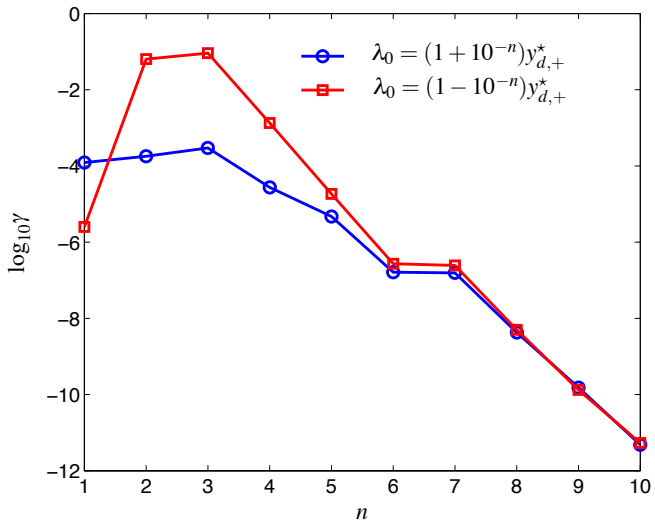


Fig. 5.13: *Mean value of the norm of H_z along the outer boundary of the bottom PML $\gamma = \langle |H_z(-\hat{h}^-)| \rangle_x$, for λ_0 approaching the Wood's anomaly $y^\star_{d,+}$ by inferior values ($\lambda_0 = (1-10^{-n})y^\star_{d,+}$, red squares) and by superior value ($\lambda_0 = (1+10^{-n})y^\star_{d,+}$, blue circles) as a function of n.*

Eventually, to illustrate the behavior of the adaptative PML when the incident wavelength gets closer to a given Wood's anomaly, we computed the mean value of the norm of H_z along the outer boundary of the bottom PML $\gamma = \langle |H_z(-\hat{h}^-)| \rangle_x$, when $\lambda_0 = (1+10^{-n})y^\star_{d,+}$ and $\lambda_0 = (1-10^{-n})y^\star_{d,+}$, for $n = 1, 2, \ldots 10$. The results are shown in Fig. 5.13. As the wavelength gets closer to $y^\star_{d,+}$, γ first increases but for $n > 3$, it decreases exponentially. However, in all cases, the value of γ remains small enough to ensure the efficiency of the PMLs.

5.2.5 *Concluding remarks*

A novel FEM formulation was adapted to the analysis of z-anisotropic gratings relying on a rigorous treatment of the plane wave sources problem through an equivalent radiation problem with localized sources. The developed approach presents the advantage of being very general in the sense that it is applicable to every conceivable grating geometry.

Numerical experiments based on existing materials at normal and oblique incidences in both TE and TM cases showed the efficiency and the accuracy of our method. We demonstrated we could generate strongly imbalanced symmetric propagative orders in the TE polarization case and at normal incidence with an aragonite grating on a silica substratum.

We also introduced the adaptative PML for grazing incidences configurations. It based on a complex-valued coordinate stretching that deals with grazing diffracted orders, yielding an efficient absorption of the field inside the PML. We provided an example in the TM polarization case (but similar results hold for the TE case), illustrating the efficiency of our method. The value of the magnetic field on the outward boundary of the PML remains small enough to consider there is no spurious reflection. The formulation is used with the FEM but can be applied to others numerical methods. Moreover, the generalization to the vectorial three-dimensional case is straightforward: the recipes given in this last section do work irrespective of the dimension and whether the problem is vectorial.

In the next section, the scalar formulation adapted to mono-dimensional gratings is extended to the the most general case of bi-dimensional grating embedded in an arbitrary multilayered dielectric stack with arbitrary incidence.

5.3 Diffraction by arbitrary crossed-gratings : a vector Finite Element formulation

5.3.1 Introduction

In this section, we extend the method detailed in Sec. 5.2 to the most general case of vector diffraction by an arbitrary crossed gratings. The main advantage of the Finite Element Method lies in its native ability to handle unstructured meshes, resulting in a build-in accurate discretization of oblique edges. Consequently, our approach remains independent of the shape of the diffractive element, whereas other methods require heavy adjustments depending on whether the geometry of the groove region presents oblique edges (*e.g.* RCWA [28], FDTD...). In this section, for the sake of clarity, we recall again the rigorous procedure allowing to deal with the issue of the plane wave sources through an equivalence of the diffraction problem with a radiation one whose sources are localized inside the diffractive element itself, as already proposed in Sec. 5.2 [29, 30].

This approach combined with the use of second order edge elements allowed us to retrieve with a good accuracy the few numerical academic examples found in the literature. Furthermore, we provide a new reference case combining major difficulties such as a non trivial toroidal geometry together with strong losses and a high permittivity contrast. Finally, we discuss computation time and convergence as a function of the mesh refinement as well as the choice of the direct solver.

5.3.2 Theoretical developments

5.3.2.1 Set up of the problem and notations

We denote by $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ the unit vectors of the axes of an orthogonal coordinate system $Oxyz$. We only deal with time-harmonic fields; consequently, electric and magnetic fields are represented by the complex vector fields $\mathbf{E}$ and $\mathbf{H}$, with a time dependance in $\exp(-i\omega t)$. Note that incident light is now propagating along the z-axis, whereas y-axis was used in the 2D case.

Besides, in this section, for the sake of simplicity, the materials are assumed to be isotropic and therefore are optically characterized by their relative permittivity ε and relative permeability μ (note that the inverse of relative permeabilities are denoted here ν). It is of importance to note that lossy materials can be studied, the relative permittivity and relative permeability being represented by complex valued functions. The crossed-gratings we are dealing with can be split into the following regions as suggested in Fig. 5.14:

- *The superstrate* ($z > z_0$) is supposed to be homogeneous, isotropic and lossless, and therefore characterized by its relative permittivity ε^+ and its relative permeability $\mu^+ (= 1/\nu^+)$ and we denote $k^+ := k_0 \sqrt{\varepsilon^+ \mu^+}$, where $k_0 := \omega/c$,

- *The multilayered stack* ($z_N < z < z_0$) is made of N layers which are supposed to be homogeneous and isotropic, and therefore characterized by their relative permittivity ε^n, their relative permeability $\mu^n (= 1/\nu^n)$ and their thickness e_n. We denote $k_n := k_0 \sqrt{\varepsilon^n \mu^n}$ for n integer between 1 and N.

- *The groove region* ($z_g < z < z_{g-1}$), which is embedded in the layer indexed g (ε^g, μ^g) of the previously described domain, is heterogeneous. Moreover the method does work irrespective of whether the diffractive elements are homogeneous: The permittivity and permeability can vary continuously (gradient index gratings) or discontinuously (step

index gratings). This region is thus characterized by the scalar fields $\varepsilon^{g'}(x,y,z)$ and $\mu^{g'}(x,y,z)(=1/\nu^{g'}(x,y,z))$. The groove periodicity along the x–axis, respectively (resp.) y–axis, is denoted d_x, resp. d_y, in the sequel.

- *The substrate* ($z < z_N$) is supposed to be homogeneous and isotropic and therefore characterized by its relative permittivity ε^- and its relative permeability $\mu^-(=1/\nu^-)$ and we denote $k^- := k_0\sqrt{\varepsilon^-\mu^-}$,

Let us emphasize the fact that the method principles remain unchanged in the case of several diffractive patterns made of distinct geometry and/or material.

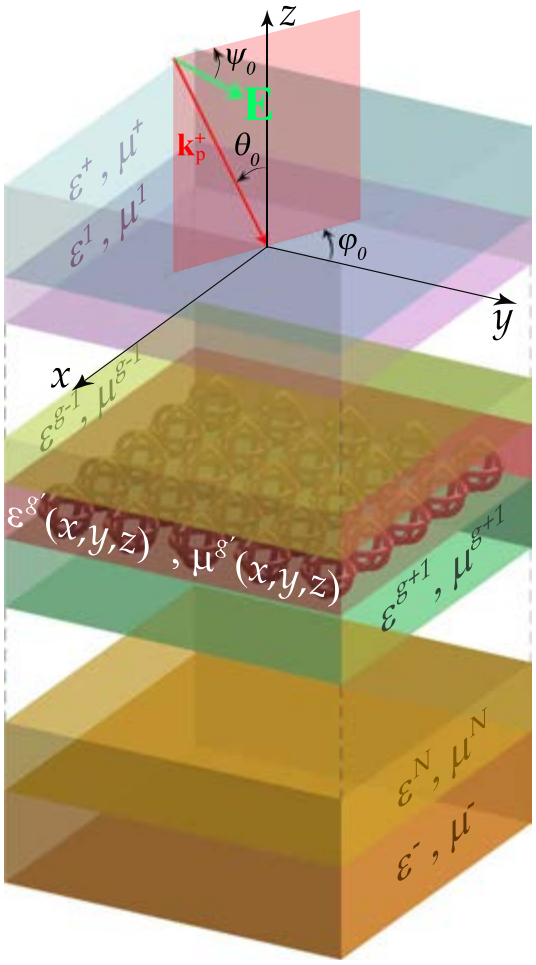


Fig. 5.14: *Scheme and notations of the studied bi-gratings.*

The incident field on this structure is denoted:

$$\mathbf{E}^{\text{inc}} = \mathbf{A}_0^e \exp(i\,\mathbf{k}_p^+ \cdot \mathbf{r}) \tag{5.53}$$

with

$$\mathbf{k}^+ = \begin{bmatrix} \alpha_0 \\ \beta_0 \\ \gamma_0 \end{bmatrix} = k^+ \begin{bmatrix} -\sin\theta_0\cos\varphi_0 \\ -\sin\theta_0\sin\varphi_0 \\ -\cos\theta_0 \end{bmatrix} \tag{5.54}$$

and

$$\mathbf{A}_0^e = \begin{bmatrix} E_x^0 \\ E_y^0 \\ E_z^0 \end{bmatrix} = A^e \begin{bmatrix} \cos\psi_0\cos\theta_0\cos\varphi_0 - \sin\psi_0\sin\varphi_0 \\ \cos\psi_0\cos\theta_0\sin\varphi_0 + \sin\psi_0\cos\varphi_0 \\ -\cos\psi_0\sin\theta_0 \end{bmatrix}, \tag{5.55}$$

where $\varphi_0 \in [0, 2\pi]$, $\theta_0 \in [0, \pi/2]$ and $\psi_0 \in [0, \pi]$ (polarization angle).

We recall here the diffraction problem: finding the solution of Maxwell equations in harmonic regime *i.e.* the unique solution $(\mathbf{E}, \mathbf{H})$ of:

$$\begin{cases} \mathbf{curl}\,\mathbf{E} = i\,\omega\,\mu_0\,\mu\,\mathbf{H} & \text{(5.56a)} \\ \mathbf{curl}\,\mathbf{H} = -i\,\omega\,\varepsilon_0\,\varepsilon\,\mathbf{E} & \text{(5.56b)} \end{cases}$$

such that the diffracted field satisfies the so-called *Outgoing Waves Condition* (OWC [31]) and where $\mathbf{E}$ and $\mathbf{H}$ are quasi-bi-periodic functions with respect to x and y coordinates.

One can choose to calculate arbitrarily $\mathbf{E}$, since $\mathbf{H}$ can be deduced from Eq. (5.56a). The diffraction problem amounts to looking for the unique solution $\mathbf{E}$ of the so-called vectorial Helmholtz propagation equation, deduced from Eqs. (5.56a,5.56b):

$$\mathscr{M}_{\varepsilon,\nu} := -\,\mathbf{curl}\,(\nu\,\mathbf{curl}\,\mathbf{E}) + k_0^2\,\varepsilon\,\mathbf{E} = \mathbf{0} \tag{5.57}$$

such that the diffracted field satisfies an OWC and where $\mathbf{E}$ is a quasi-bi-periodic function with respect to x and y coordinates.

5.3.2.2 From a diffraction problem to a radiative one with localized sources

According to Fig. 5.14, the scalar relative permittivity ε and inverse permeability ν fields associated to the studied diffractive structure can be written using complex-valued functions defined by part and taking into account the notations adopted in Sec. 5.3.2.1:

$$\upsilon(x,y,z) := \begin{cases} \upsilon^+ & \text{for} \quad z > z_0 \\ \upsilon^n & \text{for} \quad z_{n-1} > z > z_n \quad \text{with} \quad 1 \le n < g \\ \upsilon^{g'}(x,y,z) & \text{for} \quad z_{g-1} > z > z_g \\ \upsilon^n & \text{for} \quad z_{n-1} > z > z_n \quad \text{with} \quad g < n \le N \\ \upsilon^- & \text{for} \quad z < z_N \end{cases} \tag{5.58}$$

with $\upsilon = \{\varepsilon, \nu\}$, $z_0 = 0$ and $z_n = -\sum_{l=1}^{n} e_l$ for $1 \le n \le N$.
It is now convenient to introduce two functions defined by part ε_1 and ν_1 corresponding to the associated multilayered case (*i.e.* the same stack without any diffractive element) constant over Ox and Oy:

$$\upsilon_1(x,y,z) := \begin{cases} \upsilon^+ & \text{for} \quad z > 0 \\ \upsilon^n & \text{for} \quad z_{n-1} > z > z_n \quad \text{with} \quad 1 \le n \le N \\ \upsilon^- & \text{for} \quad z < z_N \end{cases} \tag{5.59}$$

with $\upsilon = \{\varepsilon, \nu\}$.
We denote by $\mathbf{E}_0$ the restriction of $\mathbf{E}^{\text{inc}}$ to the superstrate region:

$$\mathbf{E}_0 := \begin{cases} \mathbf{E}^{\text{inc}} & \text{for} \quad z > z_0 \\ \mathbf{0} & \text{for} \quad z \le z_0 \end{cases} \tag{5.60}$$

We are now in a position to define more explicitly the vector diffraction problem that we are dealing with in this section. It amounts to looking for the unique vector field $\mathbf{E}$ solution of:

$$\mathscr{M}_{\varepsilon,\nu}(\mathbf{E}) = \mathbf{0} \quad \text{such that } \mathbf{E}^d := \mathbf{E} - \mathbf{E}_0 \text{ satisfies an OWC.} \tag{5.61}$$

In order to reduce this diffraction problem to a radiation one, an intermediary vector field denoted $\mathbf{E}_1$ is necessary and is defined as the unique solution of:

$$\mathscr{M}_{\varepsilon_1,\nu_1}(\mathbf{E}_1) = \mathbf{0} \quad \text{such that } \mathbf{E}_1^d := \mathbf{E}_1 - \mathbf{E}_0 \text{ satisfies an OWC.} \tag{5.62}$$

The vector field $\mathbf{E}_1$ corresponds to an *ancillary problem* associated to the *general vectorial case of a multilayered stack* which can be calculated *independently*. This general calculation is seldom treated in the literature, we present a development in Appendix. Thus $\mathbf{E}_1$ is from now on *considered as a known* vector field. It is now apropos to introduce the unknown vector field $\mathbf{E}_2^d$, simply defined as the difference between $\mathbf{E}$ and $\mathbf{E}_1$, which can finally be calculated thanks to the FEM and:

$$\mathbf{E}_2^d := \mathbf{E} - \mathbf{E}_1 = \mathbf{E}^d - \mathbf{E}_1^d \,. \tag{5.63}$$

It is of importance to note that the presence of the superscript d is not fortuitous: As a difference between two diffracted fields (Eq. (5.63), $\mathbf{E}_2^d$ satisfies an OWC which is of prime importance in our formulation. By taking into account these new definitions, Eq. (5.61) can be written:

$$\mathscr{M}_{\varepsilon,\nu}(\mathbf{E}_2^d) = -\mathscr{M}_{\varepsilon,\nu}(\mathbf{E}_1) \,, \tag{5.64}$$

where the right-hand member is a vector field which can be interpreted as *a known vectorial source term* $-\mathscr{S}_1(x,y,z)$ *whose support is localized inside the diffractive element itself.* To prove it, let us introduce the null term defined in Eq. (5.62) and make the use of the linearity of $\mathscr{M}$, which leads to:

$$\mathscr{S}_1 := \mathscr{M}_{\varepsilon,\nu}(\mathbf{E}_1) = \mathscr{M}_{\varepsilon,\nu}(\mathbf{E}_1) - \underbrace{\mathscr{M}_{\varepsilon_1,\nu_1}(\mathbf{E}_1)}_{=\mathbf{0}} = \mathscr{M}_{\varepsilon-\varepsilon_1,\nu-\nu_1}(\mathbf{E}_1) \,. \tag{5.65}$$

5.3.2.3 Quasi-periodicity and weak formulation

The weak form is obtained by multiplying scalarly Eq. (5.61) by weighted vectors $\mathbf{E}'$ chosen among the ensemble of quasi-bi-periodic vector fields of $L^2(\mathbf{curl})$ (denoted $L^2(\mathbf{curl},(d_x,d_y),\mathbf{k})$) in Ω:

$$\mathscr{R}_{\varepsilon,\nu}(\mathbf{E},\mathbf{E}') = \int_\Omega -\mathbf{curl}\,(\nu\,\mathbf{curl}\,\mathbf{E})\cdot\overline{\mathbf{E}'} + k_0^2\,\varepsilon\,\mathbf{E}\cdot\overline{\mathbf{E}'}\,\mathrm{d}\Omega \tag{5.66}$$

Integrating by part Eq. (5.66) and making the use of the Green-Ostrogradsky theorem lead to:

$$\mathscr{R}_{\varepsilon,\nu}(\mathbf{E},\mathbf{E}') = \int_\Omega -\nu\,\mathbf{curl}\,\mathbf{E}\cdot\mathbf{curl}\,\overline{\mathbf{E}'} + k_0^2\,\varepsilon\,\mathbf{E}\cdot\overline{\mathbf{E}'}\,\mathrm{d}\Omega - \int_{\partial\Omega}(\mathbf{n}\times(\nu\,\mathbf{curl}\,\mathbf{E}))\cdot\overline{\mathbf{E}'}\,\mathrm{d}S \tag{5.67}$$

where $\mathbf{n}$ refers to the exterior unit vector normal to the surface $\partial\Omega$ enclosing Ω.

The first term of this sum concerns the volume behavior of the unknown vector field whereas the right-hand term can be used to set boundary conditions (Dirichlet, Neumann or so called quasi-periodic Bloch-Floquet conditions).

The solution $\mathbf{E}_2^d$ of the *weak form associated to the diffraction problem*, expressed in its previously defined *equivalent radiative form* at Eq. (5.64), is the element of $L^2(\mathbf{curl},(d_x,d_y),\mathbf{k})$ such that:

$$\forall\mathbf{E}' \in L^2(\mathbf{curl},d_x,d_y,\mathbf{k}),\, \mathscr{R}_{\varepsilon,\nu}(\mathbf{E}_2^d,\mathbf{E}') = -\mathscr{R}_{\varepsilon-\varepsilon_1,\nu-\nu_1}(\mathbf{E}_1,\mathbf{E}') \,. \tag{5.68}$$

In order to rigorously truncate the computation a set of Bloch boundary conditions are imposed on the pair of planes defined by $(y=-d_y/2, y=d_y/2)$ and $(x=-d_x/2, x=d_x/2)$. One can refer to [11] for a detailed implementation of Bloch conditions adapted to the FEM. A set of Perfectly Matched Layers are used in order to truncate the substrate and the superstrate along z axis (see [32] for practical implementation of PML adapted to the FEM). Since the proposed unknown $\mathbf{E}_2^d$ is quasi-bi-periodic and satisfies an OWC, this set of boundary conditions is perfectly reasonable: $\mathbf{E}_2^d$ is radiated from the diffractive element towards the infinite regions of the problem and decays exponentially inside the PMLs along z axis. The total field associated to the diffraction problem $\mathbf{E}$ is deduced at once from Eq. (5.63).

5.3.2.4 Edge or Whitney 1-form second order elements

In the vectorial case, edge elements (or Whitney forms) make a much more relevant choice [33] than nodal elements. Note that a lot of work (see for instance [34]) has been done on higher order edge elements since their introduction by Bossavit [35]. These elements are suitable to the representation of vector fields such as $\mathbf{E}_2^d$, by letting their normal component be discontinuous and imposing the continuity of their tangential components. Instead of linking the Degrees Of Freedom (DOF) of the final algebraic system to the nodes of the mesh, the DOF associated to edges (resp. faces) elements are the *circulations* (resp. *flux*) of the unknown vector field along (resp. across) its *edges* (resp. *faces*).

Let us consider the computation cell Ω together with its exterior boundary $\partial\Omega$. This volume is sampled in a finite number of tetrahedron according to the following rules: Two distinct tetrahedrons have to either share a node, an edge or a face or have no contact. Let us denote by $\mathscr{T}$ the set of tetrahedrons, $\mathscr{F}$ the set of faces, $\mathscr{E}$ the set of edges and $\mathscr{N}$ the set of nodes. In the sequel, one will refers to the node $n=\{i\}$, the edge $e=\{i,j\}$, the face $f=\{i,j,k\}$ and the tetrahedron $t=\{i,j,k,l\}$.

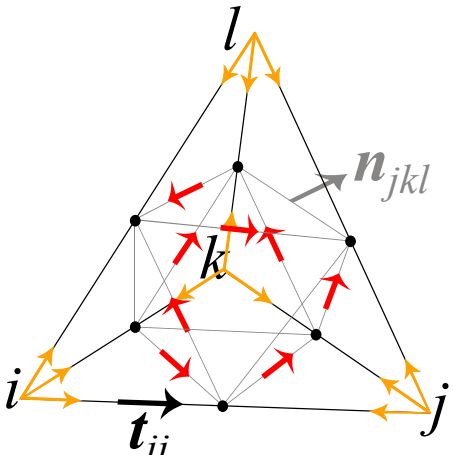


Fig. 5.15: *Degrees of freedom of a second order tetrahedral element.*

Twelve DOF (two for each of the six edges of a tetrahedron) are classically derived from line integral of weighted projection of the field $\mathbf{E}_2^d$ on each oriented edge $e=\{i,j\}$:

$$\begin{cases} \vartheta_{ij} = \displaystyle\int_i^j \mathbf{E}_2^d \cdot \mathbf{t}_{ij}\,\lambda_i\,\mathrm{d}l \\ \vartheta_{ji} = \displaystyle\int_j^i \mathbf{E}_2^d \cdot \mathbf{t}_{ji}\,\lambda_j\,\mathrm{d}l \end{cases}, \tag{5.69}$$

where $\mathbf{t}_{ij}$ is the unit vector and λ_i, the barycentric coordinate of node i, is the chosen weight function.

According to Yioultsis *et al.* [36], a judicious choice for the remaining DOF is to make the use of a tangential projection of the 1-form $\mathbf{E}_2^d$ on the face $f=\{i,j,k\}$.

$$\begin{cases} \vartheta_{ijk} = \displaystyle\int_f \left(\mathbf{E}_2^d \times \mathbf{n}_{ijk}^+\right) \cdot \mathbf{grad}\,\lambda_j\,\mathrm{d}s \\ \vartheta_{ikj} = \displaystyle\int_f \left(\mathbf{E}_2^d \times \mathbf{n}_{ijk}^-\right) \cdot \mathbf{grad}\,\lambda_k\,\mathrm{d}s \end{cases}. \tag{5.70}$$

The expressions for the shape functions, or basis vectors, of the second order 1-form Whitney element are given by:

$$\left\{ \begin{array}{ll} \mathbf{w}_{ij} & = (8\,\lambda_i^2 - 4\,\lambda_i)\,\mathbf{grad}\,\lambda_j + (-8\,\lambda_i\,\lambda_j + 2\,\lambda_j)\,\mathbf{grad}\,\lambda_i \\ \mathbf{w}_{ijk} & = 16\,\lambda_i\,\lambda_j\,\mathbf{grad}\,\lambda_k - 8\,\lambda_j\,\lambda_k\,\mathbf{grad}\,\lambda_i - 8\,\lambda_k\,\lambda_i\,\mathbf{grad}\,\lambda_j \end{array} \right. . \tag{5.71}$$

This choice of shape function ensures [37] the following fundamental property: every degree of freedom associated with a shape function should be zero for any other shape function. Finally, an approximation of the unknown $\mathbf{E}_2^d$ projected on the shape functions of the mesh m ($\mathbf{E}_2^{d,m}$) can be derived:

$$\mathbf{E}_2^{d,m} = \sum_{e\in\mathscr{E}} \vartheta_e\,\mathbf{w}_e + \sum_{f\in\mathscr{F}} \vartheta_f\,\mathbf{w}_f. \tag{5.72}$$

Weight functions $\mathbf{E}'$ (c.f. Eq. (5.68) are chosen in the same space than the unknown $\mathbf{E}_2^d$, $L^2(\mathbf{curl},(d_x,d_y),\mathbf{k})$. According to the Galerkin formulation, this choice is made so that their restriction to one bi-period belongs to the set of shape functions mentioned above. Inserting the decomposition of $\mathbf{E}_2^d$ of Eq. (5.72) in Eq. (5.68) leads to the final algebraic system which is solved, in the following numerical examples, thanks to direct solvers.

5.3.3 Energetic considerations: Diffraction efficiencies and losses

Contrarily to modal methods based on the determination of Rayleigh coefficients, the rough results of the FEM are three complex components of the vector field $\mathbf{E}^d$ interpolated over the mesh of the computation cell. Diffraction efficiencies are deduced from this field maps as follows.

As a difference between two quasi-periodic vector fields (see Eq. (5.61)), $\mathbf{E}^d$ is quasi-bi-periodic and its components can be expanded as a double Rayleigh sum:

$$E_x^d(x,y,z) = \sum_{(n,m)\in\mathbb{Z}^2} u_{n,m}^{d,x}(z)\,e^{i(\alpha_n x+\beta_m y)}, \tag{5.73}$$

with $\alpha_n = \alpha_0 + \frac{2\pi}{d_x}\,n$, $\beta_m = \beta_0 + \frac{2\pi}{d_y}\,m$ and

$$u_{n,m}^{d,x}(z) = \frac{1}{d_x\,d_y}\int_{-d_x/2}^{d_x/2}\int_{-d_y/2}^{d_y/2} E_x^d(x,y,z)\,e^{-i(\alpha_n x+\beta_m y)}\,\mathrm{d}x\,\mathrm{d}y\,. \tag{5.74}$$

By inserting the decomposition of Eq. (5.73), which is satisfied by E_x^d everywhere but in the groove region, into the Helmholtz propagation equation, one can express Rayleigh coefficients in the substrate and the superstrate as follows:

$$u_{n,m}^{d,x}(z) = e_{n,m}^{x,p}\,e^{-i\,\gamma_{n,m}^{+}\,z} + e_{n,m}^{x,c}\,e^{i\,\gamma_{n,m}^{+}\,z} \tag{5.75}$$

with $\gamma_{n,m}^{\pm^2} = k^{\pm^2} - \alpha_n^2 - \beta_m^2$, where $\gamma_{n,m}$ (or $-i\,\gamma_{n,m}$) is positive. The quantity $u_{n,m}^{d,x}$ is the sum of a propagative plane wave (which propagates towards decreasing values of z, superscript p) and of a counterpropagative one (superscript c). The OWC verified by $\mathbf{E}^d$ imposes:

$$\forall(n,m)\in\mathbb{Z}^2 \left\{ \begin{array}{lll} e_{n,m}^{x,p} = 0 & \text{for} & z > z_0 \\ e_{n,m}^{x,c} = 0 & \text{for} & z < z_N \end{array} \right. \tag{5.76}$$

Eq. (5.74) allows to evaluate numerically $e_{n,m}^{x,c}$ (resp. $e_{n,m}^{x,p}$) by double trapezoidal integration of a slice of the complex component E_x^d at an altitude z_c fixed in the superstrate (resp. substrate).

It is well known that the mere trapezoidal integration method is very efficient for smooth and periodic functions (integration on one period). The same holds for E_y^d and E_z^d components as well as their coefficients $e_{n,m}^{y,\{c,p\}}$ and $e_{n,m}^{z,\{c,p\}}$.

The dimensionless expression of the efficiency of each reflected and transmitted (n,m) order [38] is deduced from Eqs. (5.75,5.76):

$$\begin{cases} R_{n,m} = \frac{1}{|A_e|^2} \dfrac{\gamma_{n,m}^{+}}{\gamma_0} \mathbf{e}_{n,m}^{c}(z_c) \cdot \overline{\mathbf{e}_{n,m}^{c}(z_c)} & \text{for} \quad z_c > z_0 \\ T_{n,m} = \frac{1}{A_e^2} \dfrac{\gamma_{n,m}^{-}}{\gamma_0} \mathbf{e}_{n,m}^{p}(z_c) \cdot \overline{\mathbf{e}_{n,m}^{p}(z_c)} & \text{for} \quad z_c < z_N \end{cases}, \tag{5.77}$$

with $\mathbf{e}_{n,m}^{\{c,p\}} = e_{n,m}^{x,\{c,p\}}\,\mathbf{x} + e_{n,m}^{y,\{c,p\}}\,\mathbf{y} + e_{n,m}^{z,\{c,p\}}\,\mathbf{z}$.

Furthermore, normalized losses Q can be obtained through the computation of the following ratio:

$$Q = \frac{\displaystyle\int_V \frac{1}{2}\,\omega\,\varepsilon_0\,\Im m(\varepsilon^{g'})\,\mathbf{E}\cdot\overline{\mathbf{E}}\,\mathrm{d}V}{\displaystyle\int_S \frac{1}{2}\Re e\{\mathbf{E}_0 \times \overline{\mathbf{H}_0}\}\cdot\mathbf{n}\,\mathrm{d}S}. \tag{5.78}$$

The numerator in Eq. (5.78) clarifies losses in watts by bi-period of the considered crossed-grating and are computed by integrating the Joule effect losses density over the volume V of the lossy element. The denominator normalizes these losses to the incident power, *i.e.* the time-averaged incident Poynting vector flux across one bi-period (a rectangular surface S of area $d_x d_y$ in the superstrate parallel to Oxy, whose normal oriented along decreasing values of z is denoted $\mathbf{n}$). Since $\mathbf{E}_0$ is nothing but the plane wave defined at Eqs. (5.54,5.55), this last term is equal to $(A_e^2\sqrt{\varepsilon_0/\mu_0}\,d_x d_y)/(2\cos(\theta_0))$. Volumes and normal to surfaces being explicitly defined, normalized losses losses Q are quickly computed once $\mathbf{E}$ determined and interpolated between mesh nodes.

Finally, the accuracy and self-consistency of the whole calculation can be evaluated by summing the real part of transmitted and reflected efficiencies (n,m) to normalized losses:

$$Q + \sum_{(n,m)\in\mathbb{Z}^2} \Re e\{R_{n,m}\} + \sum_{(n,m)\in\mathbb{Z}^2} \Re e\{T_{n,m}\},$$

quantity to be compared to 1. The sole diffraction orders taken into account in this conservation criterium correspond to propagative orders whose efficiencies have a non-null real part. Indeed, diffraction efficiencies of evanescent orders, corresponding to pure imaginary values of $\gamma_{n,m}^{\pm}$ for higher values of (n,m) (see Eq. (5.75)) are also pure imaginary values as it appears clearly in Eq. (5.77). Numerical illustrations of such global energy balances are presented in the next section.

5.3.4 Accuracy and convergence

5.3.4.1 Classical crossed gratings

There are only a few references in the literature containing numerical examples. For each of them, the problem only consists of three regions (superstrate, grooves and substrate) as summed up on Figure 5.16. For the four selected cases, among six found in the literature, published

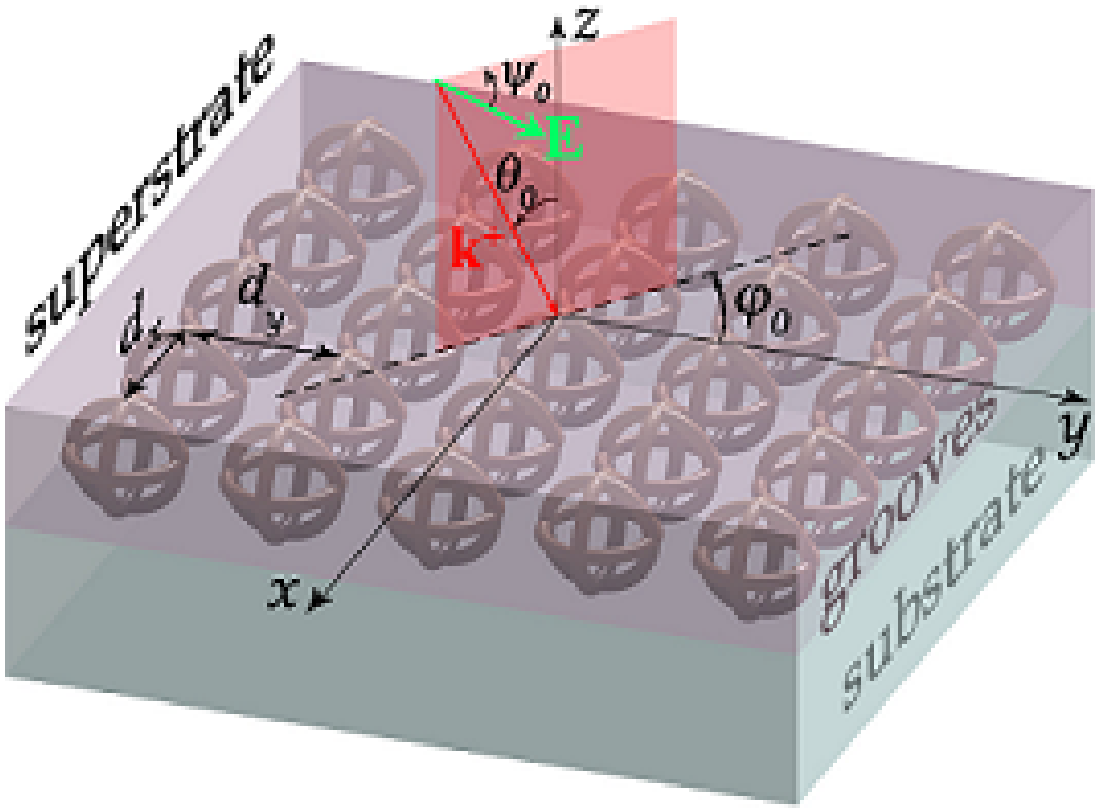
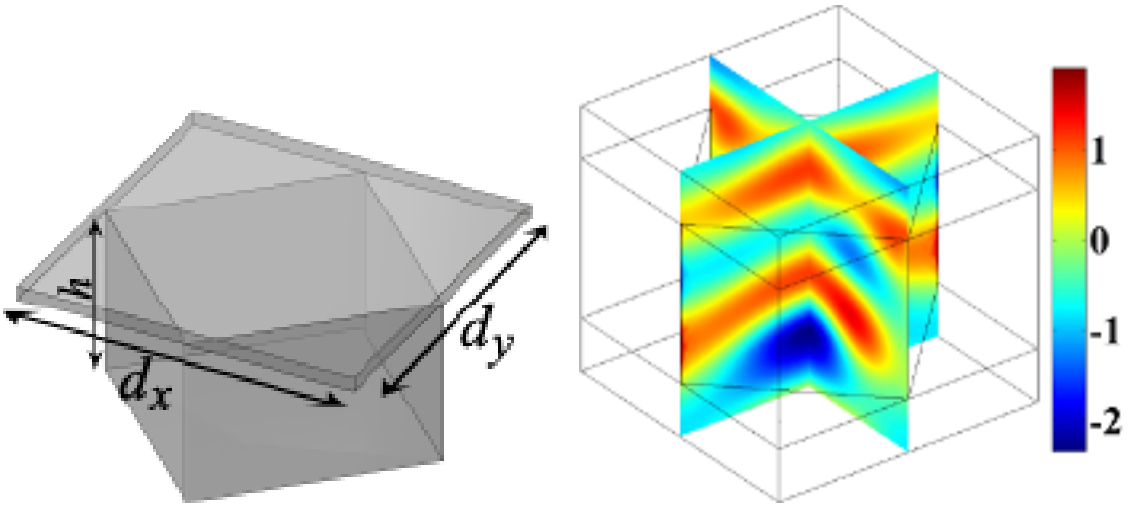


Fig. 5.16: *Configuration of the studied cases.*

results are compared to ones given by our formulation of the FEM. Moreover, in each case, a satisfying global energy balance is detailed. Finally a new validation case combining all the difficulties encountered when modeling crossed-gratings is proposed: A non-trivial geometry for the diffractive pattern (a torus), made of an arbitrary lossy material leading to a large step of index and illuminated by a plane wave with an oblique incidence. Convergence of the FEM calculation as well as computation time will be discussed in Sec. 5.3.4.2.

Checkerboard grating In this example worked out by L. Li [27], the diffractive element is a rectangular parallelepiped as shown Fig. 5.17a and the grating parameter highlighted in Fig. 5.16 are the following: $\varphi_0 = \theta_0 = 0°$, $\psi_0 = 45°$, $d_x = d_y = 5\lambda_0\sqrt{2}/4$, $h = \lambda_0$, $\varepsilon^+ = \varepsilon^{g'} = 2.25$ and $\varepsilon^- = \varepsilon^g = 1$.



(a) **(b)**

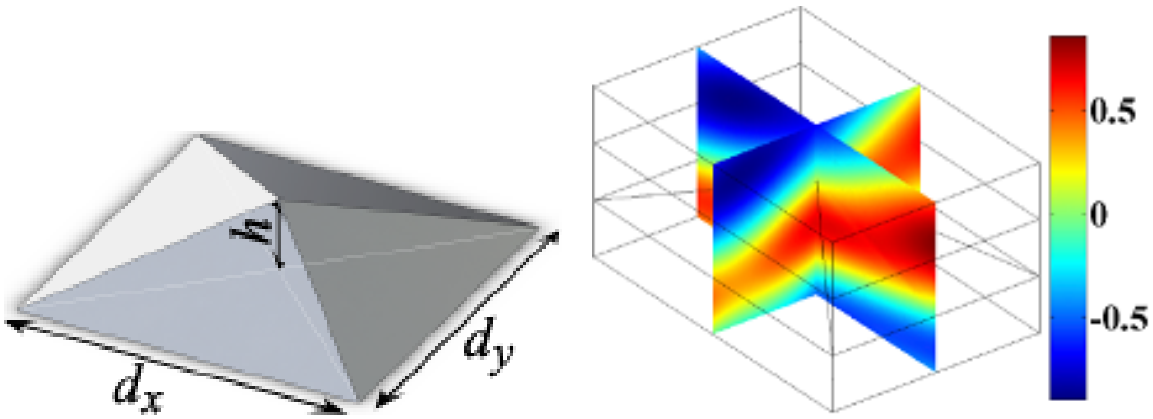
Fig. 5.17: *Diffractive element with vertical edges (a). $\Re e\{E_x\}$ in V/m (b).*

	FMM [27]	FEM
$T_{-1,-1}$	0.04308	0.04333
$T_{-1,0}$	0.12860	0.12845
$T_{-1,+1}$	0.06196	0.06176
$T_{0,-1}$	0.12860	0.12838
$T_{0,0}$	0.17486	0.17577
$T_{0,+1}$	0.12860	0.12839
$T_{+1,-1}$	0.06196	0.06177
$T_{+1,0}$	0.12860	0.12843
$T_{+1,+1}$	0.04308	0.04332
$\sum_{(n,m)\in\mathbb{Z}} \Re e\{R_{n,m}\}$	-	0.10040
TOTAL	-	1.00000

Tab. 5.4: *Energy balance [27].*

Our formulation of the FEM shows good agreement with the Fourier Modal Method developed by L. Li ([27], 1997) since the maximal relative difference between the array of values presented in Table 5.4 remains lower than 10^{-3}. Moreover, the sum of the efficiencies of propagative orders given by the FEM is very close to 1 in spite of the addition of all errors of determination upon the efficiencies.

Pyramidal crossed-grating In this example firstly worked out by Derrick *et al.* [39], the diffractive element is a pyramid with rectangular basis as shown Fig. 5.18a and the grating parameters highlighted in Fig. 5.16 are the following: $\lambda_0 = 1.533$, $\varphi_0 = 45°$, $\theta_0 = 30°$, $\psi_0 = 0°$, $d_x = 1.5$, $d_y = 1$, $h = 0.25$, $\varepsilon^+ = \varepsilon^g = 1$ and $\varepsilon^- = \varepsilon^{g'} = 2.25$. Results given by the FEM show



(a) **(b)**

Fig. 5.18: *Diffractive element with oblique edges (a). $\Re e\{E_y\}$ in V/m (b).*

Given in	[39]	[40]	[41]	[42]	FEM
$R_{-1,0}$	0.00254	0.00207	0.00246	0.00249	0.00251
$R_{0,0}$	0.01984	0.01928	0.01951	0.01963	0.01938
$T_{-1,-1}$	0.00092	0.00081	0.00086	0.00086	0.00087
$T_{0,-1}$	0.00704	0.00767	0.00679	0.00677	0.00692
$T_{-1,0}$	0.00303	0.00370	0.00294	0.00294	0.00299
$T_{0,0}$	0.96219	0.96316	0.96472	0.96448	0.96447
$T_{1,0}$	0.00299	0.00332	0.00280	0.00282	0.00290
TOTAL	0.99855	1.00001	1.00008	0.99999	1.00004

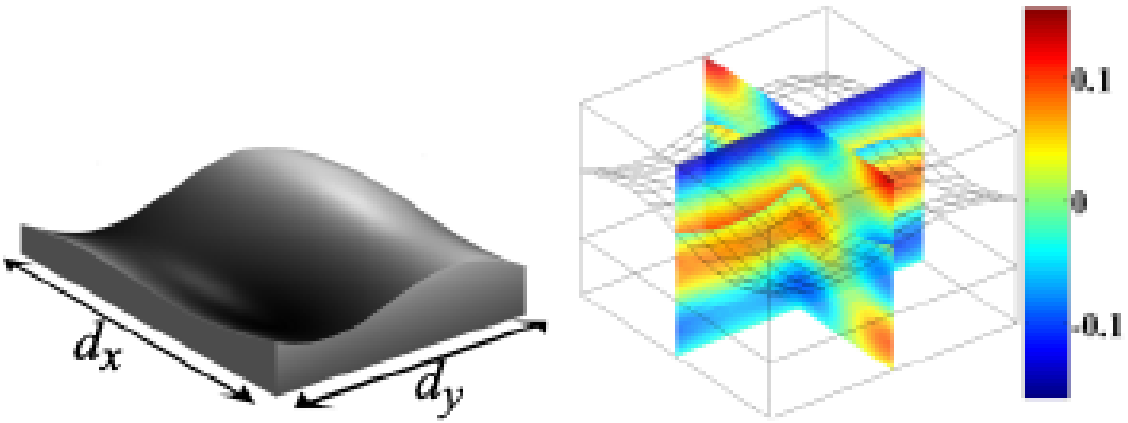
Tab. 5.5: *Comparison with the results given in [39, 40, 41, 42].*

good agreement with ones of the C method [39, 42], the Rayleigh method [40] and the RCWA [41]. Note that, in this case, some edges of the diffractive element are oblique.

Bi-sinusoidal grating In this example worked out by Bruno *et al.* [43], the surface of the grating is bi-sinusoidal (see Fig. 5.19a) and described by the function f defined by:

$$f(x,y) = \frac{h}{4}\left[\cos\left(\frac{2\pi x}{d}\right) + \cos\left(\frac{2\pi y}{d}\right)\right] \tag{5.79}$$

The grating parameters *et al.*highlighted in Fig. 5.16 are the following: $\lambda_0 = 0.83$, $\varphi_0 = \theta_0 = \psi_0 = 0°$, $d_x = d_y = 1$, $h = 0.2$, $\varepsilon^+ = \varepsilon^g = 1$ and $\varepsilon^- = \varepsilon^{g'} = 4$. Note that in order to define this



(a) **(b)**

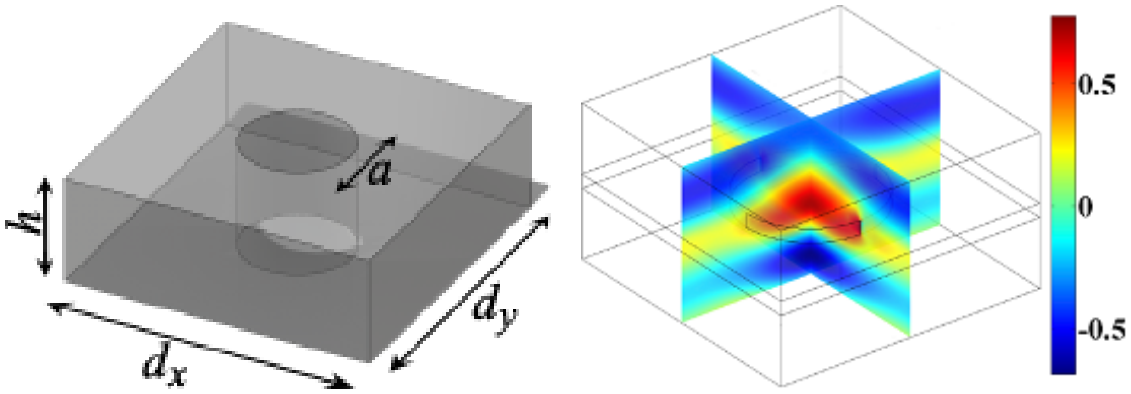
Fig. 5.19: *Diffractive element with oblique edges (a).* $\Re e\{E_z\}$ *in V/m (b).*

	[43]	FEM
$R_{-1,0}$	0.01044	0.01164
$R_{0,-1}$	0.01183	0.01165
$T_{-1,-1}$	0.06175	0.06299
$\sum_{(n,m)\in\mathbb{Z}} \Re e\{R_{n,m}\}$	-	0.10685
$\sum_{(n,m)\in\mathbb{Z}} \Re e\{T_{n,m}\}$	-	0.89121
TOTAL	-	0.99806

Tab. 5.6: *Energy balance [43].*

surface, the bi-sinusoid was first sampled (15×15 points), then converted to a 3D file format. This sampling can account for the slight differences with the results obtained using the method of variation of boundaries developed by Bruno *et al.* (1993).

Circular apertures in a lossy layer In this example worked out by Schuster *et al.* [44], the diffractive element is a circular aperture in a lossy layer as shown Fig. 5.20a and the grating parameter highlighted in Fig. 5.16 are the following: $\lambda_0 = 500\,\text{nm}$, $\varphi_0 = \theta_0 = 0°$, $\varepsilon^+ = \varepsilon^g = 1$, $\varepsilon^{g'} = 0.8125 + 5.2500\,i$ and $\varepsilon^- = 2.25$.



(a) **(b)**

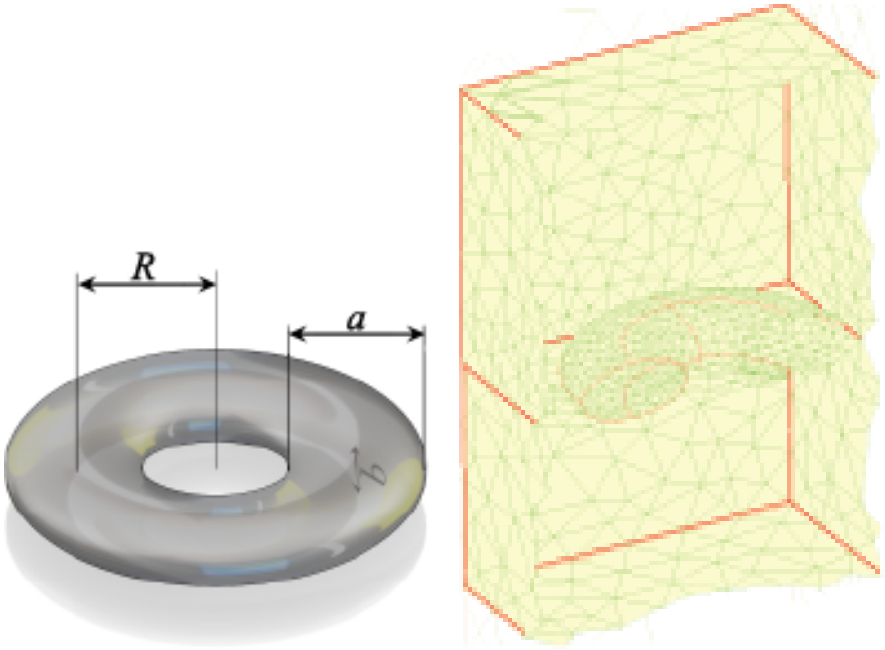
Fig. 5.20: *Lossy diffractive element with vertical edges (a). $\Re e\{E_y\}$ in V/m (b).*

	[45]	[27]	[44]	FEM
$R_{0,0}$	0.24657	0.24339	0.24420	0.24415
$\sum_{(n,m)\in\mathbb{Z}} \Re e\{T_{n,m}\}$	–	–	–	0.29110
$\sum_{(n,m)\in\mathbb{Z}} \Re e\{R_{n,m}\}$	–	–	–	0.26761
Q	–	–	–	0.44148
TOTAL	–	–	–	1.00019

Tab. 5.7: *Comparison with [45, 27, 44] and energy balance.*

In this lossy case, results obtained with the FEM show good agreement with the ones obtained with the FMM [27], the differential method [44, 46] and the RCWA [45]. Joule losses inside the diffractive element can be easily calculated, which allows to provide a global energy balance for this configuration. Finally, the convergence of the value $R_{0,0}$ as a function of the mesh refinement will be examined.

Lossy tori grating We finally propose a new test case for crossed-grating numerical methods. The major difficulty of this case lies both in the non trivial geometry (see Fig. 5.21a) of the diffractive object and in the fact that it is made of a material chosen so that losses are optimal inside it. The grating parameters highlighted in Fig. 5.16 and Fig. 5.21a are the following: $\lambda_0 = 1$, $\varphi_0 = \psi_0 = 0°$, $d_x = d_y = 0.3$, $a = 0.1$, $b = 0.05$, $R = 0.15$, $h = 500\,\text{nm}$, $\varepsilon^+ = \varepsilon^g = 1$, $\varepsilon^{g'} = -21 + 20\,i$ and $\varepsilon^- = 2.25$.



(a) **(b)**

Fig. 5.21: *Torus parameters (a). Coarse mesh of the computational domain (b).*

FEM 3D	$\theta = 0°$	$\theta = 40°$
$R_{0,0}$	0.36376	0.27331
$T_{0,0}$	0.32992	0.38191
Q	0.30639	0.34476
TOTAL	1.00007	0.99998

Tab. 5.8: *Energy balances at normal and oblique incidence.*

Tab. 5.8 illustrates the independence of our method towards the geometry of the diffractive element. $\varepsilon^{g'}$ is chosen so that the skin depth has the same order of magnitude as b, which maximizes losses. Note that energy balances remain very accurate at normal and oblique incidence, in spite of both the non-triviality of the geometry and the strong losses.

5.3.4.2 Convergence and computation time

Convergence as a function of mesh refinement When using modal methods such as the RCWA or the differential method, based on the calculation of Rayleigh coefficients, a number proportional to N_R have to be be determined *a priori*. Then, the unknown diffracted field is expanded as a Fourier serie, injected under this form in Maxwell equations, which annihilates $x-$ and $y-$dependencies. This leads to a system of coupled partial differential equations whose coefficients can structured in a matrix formalism. The resulting matrix is sometimes directly invertible (RCWA) depending on whether the geometry allows to suppress the $z-$dependance, which makes this method adapted to diffractive elements with vertically (or decomposed in staircase functions) shaped edge. In some other cases, one has to make the use of integral methods in order to solve the system, as in the pyramidal case for instance, which leads to the so-called differential method. The diffracted field map can be deduced from these coefficients. If the grating configuration only calls for a few propagative orders and if the field inside the groove region is not the main information sought for, these two close methods allow to determine the repartition of the incident energy quickly. However, if the field inside the groove region is the main piece of information, it is advisable to calculate many Rayleigh coefficients corresponding to evanescent waves which increases the computation time as $(N_R)^3$ or even $(N_R)^4$.

FEM relies on the direct calculation of the vectorial components of the complex field. Rayleigh coefficients are determined *a posteriori*. The parameter limiting the computation time is the number of tetrahedral elements along which the computational domain is split up. We suppose that it is necessary to calculate at least two or three points (or mesh nodes) per period of the field ($\lambda_0/\sqrt{\Re e\{\varepsilon\}}$). Figure 5.22 shows the convergence of the efficiency $R_{0,0}$ (circular apertures case, see Fig. 5.20a) as a function of the mesh refinement characterized by the parameter N_M: The maximum size of each element is set to $\lambda_0/(N_M\sqrt{\Re e\{\varepsilon\}})$.

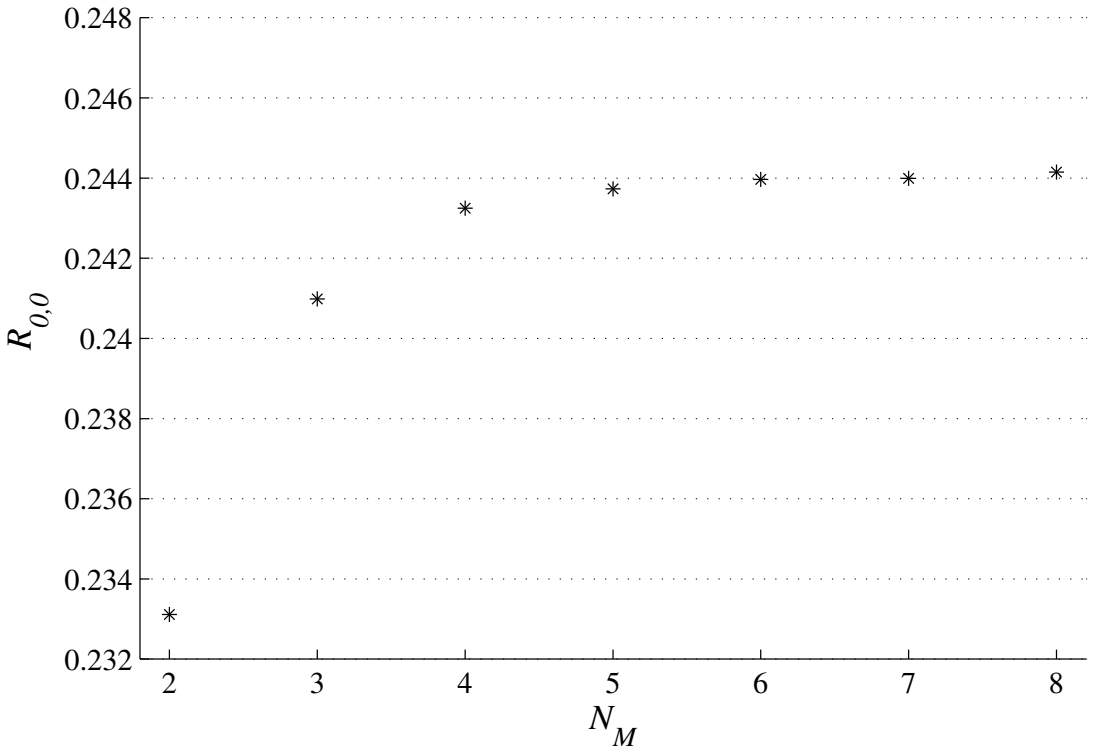


Fig. 5.22: *Convergence of $R_{0,0}$ in function of N_m (circular apertures crossed-grating).*

It is of interest to note that even if $N_M < 3$ the FEM still gives pertinent diffraction efficiencies: $R_{0,0} = 0.2334$ for $N_M = 1$ and $R_{0,0} = 0.2331$ for $N_M = 2$. The Galerkin method (see Eq. (5.67)) corresponds to a minimization of the error (between the exact solution and the approximation) with respect to a norm that can be physically interpreted in terms of energy-related quantities. Therefore, the finite element methods usually provide energy-related quantities that are more accurate than the local values of the fields themselves.

Computation time All the calculations were performed on a server equipped with 8 dual core Itanium1 processors and 256Go of RAM. Tetrahedral quadratic edge elements were used together with the direct solver PARDISO. Among different direct solvers adapted to sparse matrix algebra (UMFPACK, SPOOLES and PARDISO), PARDISO turned out to be the less time-consuming one as shown in Tab 5.9.

Solver	Computation time for 41720 DOF	Computation time for 205198 DOF
SPOOLES	15 mn 32 s	14 h 44 mn
UMFPACK	2 mn 07 s	1 h 12 mn
PARDISO	57 s	16 mn

Tab. 5.9: *Computation time variations from solver to solver.*

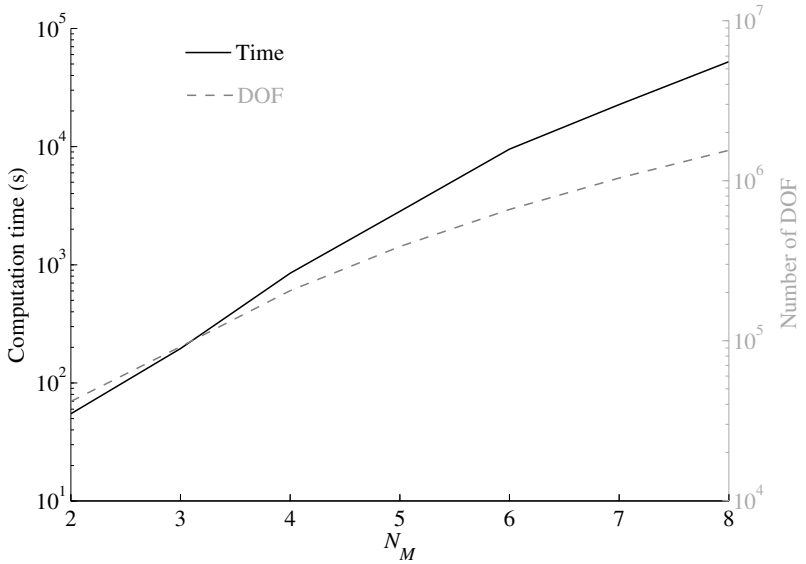


Fig. 5.23: *Computation time and number of DOF as a function of N_M.*

Figure 5.23 shows the computation time required to perform the whole FEM computational process for a system made of a number of DOF indicated on the right-hand ordinate. It is of importance to note that for values of N_M lower than 3, the problem can be solved in less than a minute on a standard laptop (4Go RAM, 2×2 GHz) with 3 significant digits on the diffraction efficiencies. This accuracy is more than sufficient in numerous experimental cases. Furthermore, as far as integrated values are at stake, relatively coarse meshes ($N_M \approx 1$) can be used trustfully, authorizing fast geometric, spectral or polarization studies.

Nowadays, the efficiency of the numerical algorithms for sparse matrix algebra together with the available power of computers and the fact that the problem reduces to a basic cell with a

size of a small number of wavelengths make the finite element problem very tractable as proved here.

5.4 Concluding remarks

In this chapter, we demonstrate a general formulation of the FEM allowing to calculate the diffraction efficiencies from the electromagnetic field diffracted by arbitrarily shaped gratings embedded in a multilayered stack lightened by a plane wave of arbitrary incidence and polarization angle. It relies on a rigorous treatment of the plane wave sources problem through an equivalent radiation problem with localized sources. Bloch conditions and a new dedicated PML have been implemented in order to rigorously truncate the computational domain.

The principles of the method were discussed in detail for mono-dimensional gratings in TE/TM polarization cases (2D or scalar case) in a first part, and for the most general bi-dimensional or crossed gratings (3D or vector case) in a second part. Note that the very same concepts could be applied to the intermediate case of mono-dimensional gratings enlighten by an arbitrary incident plane wave (so-called conical case). The reader will find detail about the element basis relevant to this case in [11].

The main advantage of this formulation is its complete generality with respect to the studied geometries and the material properties, as illustrated with the lossy tori grating non-trivial case. Its principle remains independent of both the number of diffractive elements by period and number of stack layers. Its flexibility allowed us to retrieve with accuracy the few numerical academic examples found in the literature and established with independent methods.

The remarkable accuracy observed in the case of coarse meshes, makes it a fast tool for the design and optimization of diffractive optical components (*e.g.* reflection and transmission filters, polarizers, beam shapers, pulse compression gratings...). The complete independence of the presented approach towards both the geometry and the isotropic constituent materials of the diffractive elements makes it a handy and powerful tool for the study of metamaterials, finite-size photonic crystals, periodic plasmonic structures...The method described in this chapter has already been successfully applied to various problems, from homogenization theory [47] or transformation optics [48] to more applied concerns as the modeling of complex CMOS nanophotonic devices [49] or ultra-thin new generation solar cells [50].

5.A APPENDIX

This appendix is dedicated to the determination of the vector electric field in a dielectric stack enlightened by a plane wave of arbitrary polarization and incidence angle. This calculation, abundantly treated in the 2D scalar case, is generally not presented in the literature since, as far as isotropic cases are concerned, it is possible to project the general vectorial case on the two reference TE and TM cases. However, the presented formulation can be extended to a fully anisotropic case for which this TE/TM decoupling is no longer valid and the three components of the field have to be calculated as follows.

Let us consider the *ancillary problem* mentioned in Sec. 5.3.2.2, *i.e.* a dielectric stack made of N homogeneous, isotropic, lossy layers characterized by there relative permittivity denoted ε^j and their thickness e_j. This stack is deposited on a homogeneous, isotropic, possibly lossy substrate characterized by its relative permittivity denoted $\varepsilon^{N+1} = \varepsilon^-$. The superstrate is air and its relative permittivity is denoted $\varepsilon^+ = 1$. Finally, we denote by z_j the altitude of the interface between the j^{th} and $j+1^{th}$ layers. The restriction of the incident field $\mathbf{E}^{\text{inc}}$ to the superstrate region is denoted $\mathbf{E}_0$. The problem amounts to looking for ($\mathbf{E}_1$,$\mathbf{H}_1$) satisfying Maxwell equations in harmonic regime (see Eqs. (5.56a,5.56b)).

Across the interface $z = z_j$

By projection on the main axis of the vectorial Helmholtz propagation equation (Eq. (5.57)), the total electric field inside the j^{th} layer can be written as the sum of a propagative and a counter-propagative plane waves:

$$\mathbf{E}_1(x,y,z) = \begin{bmatrix} E_1^{x,j,+} \\ E_1^{y,j,+} \\ E_1^{z,j,+} \end{bmatrix} \exp\left(j\left(\alpha_0 x + \beta_0 y + \gamma_j z\right)\right) + \begin{bmatrix} E_1^{x,j,-} \\ E_1^{y,j,-} \\ E_1^{z,j,-} \end{bmatrix} \exp\left(j\left(\alpha_0 x + \beta_0 y - \gamma_j z\right)\right) \tag{5.80}$$

where

$$\gamma_j^2 = k_j^2 - \alpha_0^2 - \beta_0^2 \tag{5.81}$$

What follows consists in writing the continuity of the tangential components of $(\mathbf{E}_1, \mathbf{H}_1)$ across the interface $z = z_j$, *i.e.* the continuity of the vector field Ψ defined by:

$$\Psi = \begin{bmatrix} E_1^x \\ E_1^y \\ iH_1^x \\ iH_1^y \end{bmatrix}. \tag{5.82}$$

The continuity of Ψ along Oz together with its analytical expression inside the j^{th} and $j+1^{th}$ layers allows to establish a recurrence relation for the interface $z = z_j$.

Then, by projection of Eqs. (5.56a,5.56b) on Ox,Oy and Oz:

$$\begin{bmatrix} i\beta_0 H_1^z - \frac{\partial H_1^y}{dz} \\ \frac{\partial H_1^x}{dz} - i\alpha_0 H_1^z \\ i\alpha_0 H_1^y - i\beta_0 H_1^x \end{bmatrix} = -i\omega\varepsilon \begin{bmatrix} E_1^x \\ E_1^y \\ E_1^z \end{bmatrix} \tag{5.83}$$

and

$$\begin{bmatrix} i\beta_0 E_1^z - \frac{\partial E_1^y}{\partial z} \\ \frac{\partial E_1^x}{\partial z} - i\alpha_0 E_1^z \\ i\alpha_0 E_1^y - i\beta_0 E_1^x \end{bmatrix} = i\omega\mu \begin{bmatrix} H_1^x \\ H_1^y \\ H_1^z \end{bmatrix}. \tag{5.84}$$

Consequently, tangential components of $\mathbf{H}_1$ can be expressed in function of tangential components of $\mathbf{E}_1$:

$$\underbrace{\begin{bmatrix} \omega\mu & 0 & \beta_0 \\ 0 & \omega\mu & -\alpha_0 \\ -\beta_0 & \alpha_0 & -\omega\varepsilon \end{bmatrix}}_{B} \begin{bmatrix} iH_1^x \\ iH_1^y \\ iH_1^z \end{bmatrix} = \begin{bmatrix} \frac{\partial E_1^y}{dz} \\ -\frac{\partial E_1^x}{dz} \\ 0 \end{bmatrix}. \tag{5.85}$$

By noticing the invariance and linearity of the problem along Ox and Oy, the following notations are adopted:

$$\begin{cases} U_x^{j,\pm} = E_1^{x,j,\pm} \exp(\pm i\gamma_j z) \\ U_y^{j,\pm} = E_1^{y,j,\pm} \exp(\pm i\gamma_j z) \end{cases} \tag{5.86}$$

and

$$\Phi_j = \begin{bmatrix} U_x^{+,j} \\ U_x^{-,j} \\ U_y^{+,j} \\ U_y^{-,j} \end{bmatrix}. \tag{5.87}$$

Thanks to Eq. (5.80) and Eq. (5.84) and letting $M = B^{-1}$, it comes for the j^{th} layer:

$$\Psi(x,y,z) = \exp(i(\alpha_0 x + \beta_0 y)) \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \gamma_j M_{12}^j & -\gamma_j M_{12}^j & -\gamma_j M_{11}^j & \gamma_j M_{11}^j \\ \gamma_j M_{22}^j & -\gamma_j M_{22}^j & -\gamma_j M_{21}^j & \gamma_j M_{21}^j \end{bmatrix}}_{\Pi_j} \begin{bmatrix} U_x^{+,j} \\ U_x^{-,j} \\ U_y^{+,j} \\ U_y^{-,j} \end{bmatrix}. \tag{5.88}$$

Finally, the continuity of Ψ at the interface $z = z_j$ leads to:

$$\Phi_{j+1}(z_j) = \Pi_{j+1}^{-1} \Pi_j \Phi_j(z_j). \tag{5.89}$$

Normal components can be deduced using Eqs. (5.83,5.84).

Traveling inside the $j+1^{th}$ layer

Using Eq. (5.80), a simple phase shift allows to travel from $z = z_j$ to $z = z_{j+1} = z_j - e_{j+1}$:

$$\Phi_{j+1}(z_{j+1}) = \underbrace{\begin{bmatrix} \exp(-i\gamma_{j+1} e_{j+1}) & 0 & 0 & 0 \\ 0 & \exp(+i\gamma_{j+1} e_{j+1}) & 0 & 0 \\ 0 & 0 & \exp(-i\gamma_{j+1} e_{j+1}) & 0 \\ 0 & 0 & 0 & \exp(+i\gamma_{j+1} e_{j+1}) \end{bmatrix}}_{T_{j+1}} \Phi_{j+1}(z_j) \tag{5.90}$$

Thanks to Eq. (5.90) and Eq. (5.89), a recurrence relation can be formulated for the analytical expression of $\mathbf{E}_1$ in each layer:

$$\Phi_{j+1}(z_{j+1}) = T_{j+1} \Pi_{j+1}^{-1} \Pi_j \Phi_j(z_j) \tag{5.91}$$

Reflection and transmission coefficients

The last step consists in the determination of the first term Φ_0, which is not entirely known, since the problem definition only specifies $U_x^{0,+}$ and $U_y^{0,+}$, imposed by the incident field $\mathbf{E}_0$. Let us make the use of the OWC hypothesis verified by $\mathbf{E}_1^d$ (see Eq. (5.62)). This hypothesis directly translates the fact that none of the components of $\mathbf{E}_1^d$ can either be traveling down in the superstrate or up in the substrate: $U_y^{N+1,-} = U_x^{N+1,-} = 0$. Therefore, the four unknowns $U_x^{0,-}$, $U_y^{0,-}$,$U_y^{N+1,+}$ and $U_x^{N+1,+}$, *i.e.* transverse components of the vector fields reflected and transmitted by the stack, verify the following equation system:

$$\Phi_{N+1}(z_N) = (\Pi_{N+1})^{-1}\,\Pi_N \prod_{j=0}^{N-1} T_{N-j}\,(\Pi_{N-j})^{-1}\,\Pi_{N-j-1}\,\Phi_0(z_0) \tag{5.92}$$

This allows to extend the definition of transmission and reflection widely used in the scalar case. Finally, Φ_{N+1} is entirely defined. Making the use of the recurrence relation of Eq. (5.91) and of Eq. (5.80) leads to an analytical expression for $\mathbf{E}_1^d$ in each layer.

References:

[1] A. Bossavit, "Solving Maxwell equations in a closed cavity, and the question ofspurious modes," IEEE Trans. on Mag. **26**, 702–705 (1990).

[2] J-C. Nedelec, "Mixed finite elements in R 3," Numerische Mathematik **35**, 315–341 (1980).

[3] A. Bossavit, "Solving maxwell equations in a closed cavity, and the question ofspurious modes'," Magnetics, IEEE Transactions on **26**, 702–705 (1990).

[4] J-P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves," J. Comput. Phys. **114**, 185–200 (1994).

[5] W. Chew and W. Weedon, "A 3d perfectly matched medium from modified maxwell's equations with stretched coordinates," Microwave and optical technology letters **7**, 599–604 (2007).

[6] F. Teixeira and W. Chew, "General closed-form pml constitutive tensors to match arbitrary bianisotropic and dispersive linear media," Microwave and Guided Wave Letters, IEEE **8**, 223–225 (1998).

[7] A. Nicolet, F. Zolla, Y. Agha, and S. Guenneau, "Geometrical transformations and equivalent materials in computational electromagnetism," COMPEL **27**, 806–819 (2008).

[8] Y. Agha, F. Zolla, A. Nicolet, and S. Guenneau, "On the use of pml for the computation of leaky modes: An application to microstructured optical fibres," COMPEL **27**, 95–109 (2008).

[9] R. Petit, L. Botten *et al.*, *Electromagnetic theory of gratings*, vol. 62 (Springer-Verlag Berlin, 1980).

[10] F. Zolla, G. Renversez, and A. Nicolet, *Foundations of Photonic Crystal Fibres: 2nd Edition* (Imperial College Press, 2012).

[11] A. Nicolet, S. Guenneau, C. Geuzaine and F. Zolla, "Modelling of electromagnetic waves in periodic media with finite elements," J. of Comput. and Applied Math. **168**, 321–329 (2004).

[12] A. Nicolet, F. Zolla, Y. Ould Agha and S. Guenneau, "Leaky modes in twisted microstructured optical fibres," Waves in Random and Complex Media **17**, 559–570 (2007).

[13] M. Lassas, J. Liukkonen and E. Somersalo, "Complex riemannian metric and absorbing boundary condition," Journal de Mathématiques Pures et Appliquées **80**, 739–768 (2001).

[14] J. L. M. Lassas and E. Somersalo, "Analysis of the PML equations in general convex geometry," Proceedings of the Royal Society of Edinburgh **131**, 1183–1207 (2001).

[15] A. Nicolet, F. Zolla, Y. O. Agha, and S. Guenneau, "Geometrical transformations and equivalent materials in computational electromagnetism," COMPEL **27**, 806–819 (2008).

[16] P. Helluy, S. Maire and P. Ravel, "Intégrations numériques d'ordre élevé de fonctions régulières ou singulières sur un intervalle," CR. Acad. Sci. Paris, Sér. I, Math **327**, 843–848 (1998).

[17] G. Granet, "Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution," J. Opt. Soc. Am. A **16**, 2510–2516 (1999).

[18] G. Bao, Z. Chen and H. Wu, "Adaptive finite-element method for diffraction gratings," J. Opt. Soc. Am. A **22**, 1106–1114 (2005).

[19] G. Tayeb., *Contribution à l'étude de la diffraction des ondes électromagnétiques par des réseaux. Réflexions sur les méthodes existantes et sur leur extension aux milieux anisotropes.* (Thèse de doctorat en sciences (PhD), Université Aix-Marseille III, 1990).

[20] Y. Ohkawa, Y. Tsuji and M. Koshiba, "Analysis of anisotropic dielectric grating diffraction using the finite-element method," J. Opt. Soc. Am. A **13**, 1006–1012 (1996).

[21] N. Kono and Y. Tsuji, "A novel finite-element method for nonreciprocal magneto-photonic crystal waveguides," Journal of lightwave technology **22**, 1741 (2004).

[22] A. Zhou, J. Erwin, C. Brucker, and M. Mansuripur, "Dielectric tensor characterization for magneto-optical recording media," Applied optics **31**, 6280–6286 (1992).

[23] R. W. Wood, "On a Remarkable Case of Uneven Distribution of Light in a Diffraction Grating Spectrum," Proceedings of the Physical Society of London **18**, 269–275 (1902).

[24] L. Rayleigh, "Note on the remarkable case of diffraction spectra described by Prof. Wood," Philos. Mag **14**, 60–65 (1907).

[25] Z. Chen and X. Liu, "An adaptive perfectly matched layer technique for time-harmonic scattering problems," SIAM J. Numer. Anal. **43**, 645–671 (2005).

[26] A. Schädle, L. Zschiedrich, S. Burger, R. Klose, and F. Schmidt, "Domain decomposition method for maxwell's equations: Scattering off periodic structures," Journal of Computational Physics **226**, 477–493 (2007).

[27] L. Li, "New formulation of the fourier modal method for crossed surface-relief gratings," J. Opt. Soc. Am. A **14**, 2758–2767 (1997).

[28] E. Popov, M. Nevière, B. Gralak and G. Tayeb, "Staircase approximation validity for arbitrary-shaped gratings," J. Opt. Soc. Am. A **19**, 33–42 (2002).

[29] G. Demésy, F. Zolla, A. Nicolet, M. Commandré and C. Fossati, "The finite element method as applied to the diffraction by an anisotropic grating," Optics Express **15**, 18089–18102 (2007).

[30] G. Demésy, F. Zolla, A. Nicolet, M. Commandré, C. Fossati, O. Gagliano, S. Ricq and B. Dunne, "Finite element method as applied to the study of gratings embedded in complementary metal-oxide semiconductor image sensors," Optical Engineering **48**, 058002 (2009).

[31] F. Zolla and R. Petit, "Method of fictitious sources as applied to the electromagnetic diffraction of a plane wave by a grating in conical diffraction mounts," J. Opt. Soc. Am. A **13**, 796–802 (1996).

[32] Y. Ould Agha, F. Zolla, A. Nicolet and S. Guenneau, "On the use of PML for the computation of leaky modes : an application to gradient index MOF," COMPEL **27**, 95–109 (2008).

[33] P. Dular, A. Nicolet, A. Genon and W. Legros, "A discrete sequence associated with mixed finite elements and itsgauge condition for vector potentials," IEEE Transactions on Magnetics **31**, 1356–1359 (1995).

[34] P. Ingelstrom, "A new set of H (curl)-conforming hierarchical basis functions for tetrahedral meshes," IEEE Trans. Microwave Theory Tech. **54**, 106–114 (2006).

[35] A. Bossavit and I. Mayergoyz, "Edge-elements for scattering problems," IEEE Trans. on Mag. **25**, 2816–2821 (1989).

[36] T. V. Yioultsis and T. D. Tsiboukis, "The Mystery and Magic of Whitney Elements - An Insight in their Properties and Construction," ICS Newsletter **3**, 1389–1392 (Nov. 1996).

[37] T. V. Yioultsis and T. D. Tsiboukis, "Multiparametric vector finite elements: a systematic approach to the construction of three-dimensional, higher order, tangential vector shape functions," IEEE Trans. on Mag. **32**, 1389–1392 (1996).

[38] E. Noponen and J. Turunen, "Eigenmode method for electromagnetic synthesis of diffractive elements with three-dimensional profiles," J. Opt. Soc. Am. A **11**, 2494–2502 (1994).

[39] G. H. Derrick, R. C. McPhedran, D. Maystre and M. Nevière, "Crossed gratings: A theory and its applications," Appl. Phys. B **18**, 39–52 (1979).

[40] J. J. Greffet, C. Baylard and P. Versaevel, "Diffraction of electromagnetic waves by crossed gratings: a series solution," Opt. Lett. **17**, 1740–1742 (1992).

[41] R. Bräuer and O. Bryngdahl, "Electromagnetic diffraction analysis of two-dimensional gratings," Opt. Commun. **100** (1993).

[42] G. Granet, "Analysis of diffraction by surface-relief crossed gratings with use of the Chandezon method: Application to multilayer crossed gratings," J. Opt. Soc. Am. A **15**, 1121–1131 (1998).

[43] O. P. Bruno and F. Reitich, "Numerical solution of diffraction problems: a method of variation of boundaries. III. doubly periodic gratings," J. Opt. Soc. Am. A **10**, 2551–2562 (1993).

[44] T. Schuster, J. Ruoff, N. Kerwien, S. Rafler and W. Osten, "Normal vector method for convergence improvement using the rcwa for crossed gratings," J. Opt. Soc. Am. A **24**, 2880–2890 (2007).

[45] M. G. Moharam, E. B. Grann, D. A. Pommet and T. K. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," J. Opt. Soc. Am. A **12**, 1068–1076 (1995).

[46] L. Arnaud, *Diffraction et diffusion de la lumière : modélisation tridimensionnelle et application à la métrologie de la microélectronique et aux techniques d'imagerie sélective en milieu diffusant* (PhD Thesis, Université Aix-Marseille III, 2008).

[47] A. Cabuz, A. Nicolet, F. Zolla, D. Felbacq, and G. Bouchitté, "Homogenization of nonlocal wire metamaterial via a renormalization approach," J. Opt. Soc. Am. B **28**, 1275–1282 (2011).

[48] G. Dupont, S. Guenneau, S. Enoch, G. Demesy, A. Nicolet, F. Zolla, and A. Diatta, "Revolution analysis of three-dimensional arbitrary cloaks," Optics Express **17**, 22603–22608 (2009).

[49] G. Demésy, F. Zolla, A. Nicolet, M. Commandré, C. Fossati, O. Gagliano, S. Ricq, and B. Dunne, "Finite element method as applied to the study of gratings embedded in complementary metal-oxide semiconductor image sensors," Optical Engineering **48**, 058002–058002 (2009).

[50] G. Demésy and S. John, "Solar energy trapping with modulated silicon nanowire photonic crystals," Journal of Applied Physics **112**, 074326–074326 (2012).

Chapter 6:

Spherical harmonic Lattice Sums for Gratings

Brian Stout

Table of Contents:

Chapter 6

Spherical harmonic Lattice Sums for Gratings

Brian Stout

Institut Fresnel,Marseille, France
brian.stout@fresnel.fr

6.1 Introduction to particulate gratings

Lattice sums of spherical harmonic functions are well suited for modeling gratings composed of periodic arrays of identical discrete particles, henceforth referred to as particulate gratings (cf. fig. (6.1)). By *discrete* particles, we mean that the particles have a physical boundary such that there exists a region between the individual particles that is governed by the host material's constitutive relations. This feature makes particulate gratings somewhat different from most of the other diffraction grating problems studied in this book which are usually characterized by a substrate and a superstrate with distinct constitutive parameters. The techniques of this chapter can be extended to include the effects of a nearby planar interface,[26, 27, 28] but such considerations complicate the problem somewhat and this chapter therefore concentrates on substrate-free particulate gratings.

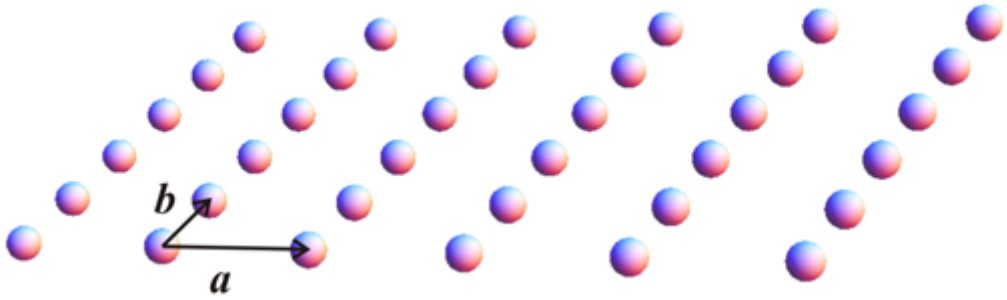


Figure 6.1: Particulate grating with lattice vectors $\mathbf{a}$ *and* $\mathbf{b}$.

Theoretical analysis of the particulate grating problem can draw on both single-particle scattering theory and techniques originally developed for solid state physics. The solid state analogy is clear from the similarity of this problem to the scattering of waves by crystal lattices, particularly in the "muffin tin" approximation[25]. Summations of the spherical harmonic fields scattered by the (infinite) number of particles in the lattice involve semi-convergent series and will generally go under the name of "lattice sums". By lattice sum, we mean sums of the form

$\sum_\Lambda \Phi(\boldsymbol{r}_j)$ where Λ refers to the ensemble of points, $\boldsymbol{r}_j$, in a periodic lattice, and Φ is a given function.

Lattice sums have applications in many fields and their study dates back to the 19th century treating conditionally convergent sums of solutions to the Laplace equations (most notably in the Madelund constant of ionic crystals). Nevertheless, they were not always recognized as a specific branch of study, and their derivations tended to be scattered throughout the literature. This situation is changing however with the appearance of extensive reviews in recent years[15, 16, 19, 20]. Furthermore, another monograph, dedicated entirely to lattice sums, was published at around the same time as this one.[3]

As developed in detail in the aforementioned monograph, the study of the (scale invariant) Laplace equation lattice sums have generated a number of important analytic results. The grating problem on the other hand involves propagating waves and consequently requires lattice sums of Helmholtz-type solutions. Although there are fewer fully analytic results for the (scale dependent) Helmholtz lattice sums than for the Laplace equation case, analytic manipulations remain essential for regularizing and accelerating the numerical analysis of Helmholtz lattice sums.

In solid-state physics, Helmholtz (i.e. Schrödinger) equation lattice sums are a key aspect of the Korringa-Kohn-Rostoker (KKR) methods for band-structure calculations in crystals.[22, 14, 13] In KKR theory, lattice sums intervene in the calculation of the "*structure constants*" of the lattice Green function and their regularization generally goes under the name of *Ewald sum* techniques. The Ewald sum method is quite intricate, but its basic principle can be viewed as separating a semi-convergent sum into slowly and rapidly convergent parts and then to transform the slowly convergent part into reciprocal space via the Poisson sum formula where it becomes a rapidly convergent series.

Although Ewald sum methods are proven to be quite efficient for most of the problems encountered in solid state physics, their utility has been repeatedly criticized for grating-type applications (requiring numerically unwieldy evaluations of incomplete Gamma functions with negative real arguments[32], poor numerical properties for high multipole orders or large wavenumber, k, *etc.*). A number of authors have consequently looked for alternative lattice sum techniques since the pioneering work of Kerker over 30 years ago. In this chapter, we simply discuss and compare some of our preferred methods in the appendices. Our emphasis will instead be placed on painting a complete gratings-picture analysis capable of describing both near and far-field phenomenon in particulate gratings.

The matrix elements of the Ω propagation matrix introduced in section 6.3.4 correspond to the "structure constants" of a KKR theory. More precisely, due to the differences between the Schrödinger and Maxwell equations, the Ω matrix elements will be shown to be written as a superposition of the KKR structure constants. In both KKR and particulate grating theory, one desires to calculate the lattice Green function. A fundamental method choice in this chapter is to use the language of T-matrices. Notably, we will see that the quasi-periodic Green function can be expressed as a lattice sum of *multiple-scattering* T-matrices. The multiple-scattering T-matrices themselves are calculated in terms of the *single-particle* T-matrices, and the Ω matrix.

The T-matrix manipulations are carried out on a basis set of solutions to the Helmholtz equation, which we generally refer to as *partial waves*, (PWs), also commonly referred to as spherical wave functions (SWFs). This T-matrix approach is also generally adopted in the KKR calculations[22], but in the light scattering community, the terminology "T-matrix method" is often considered to be synonymous with extended boundary condition technique (also called

Null-field methods), but the T-matrix is a general theoretical construct that relates the field incident on a particle to the field scattered by the particle. As such, it can be seen as providing a complete solution to the single-particle scattering problem. In practice, the T-matrix can be generated by a wide variety of techniques including DDA, method of moments, and fictitious source techniques.

The T-matrix of an individual particle depends on the shape and the constitutive parameters of the particles, both of which can be quite arbitrary as long as the particle response is linear (including anisotropic constitutive parameters, magnetic permeability contrast etc.). However, since the T-matrix can be viewed as being the complete solution of a 1-body problem, its determination can be viewed as being separate from the grating problem. In this chapter, we simplify the T-matrix part of the problem by considering only isotropic spherical scatterers. The T-matrix of such scatterers is then diagonal in the partial wave basis with its elements being determined analytically from Mie theory (cf. section 6.3.2). We insist however, that for particulate gratings composed of more exotic scatterers like split rings, it generally suffices to insert the appropriate T-matrix to obtain the response of lattices composed of such scatterers. We refer the interested to reader to reviews of the T-matrix methods.[17, 18]

The methods developed in this chapter can be adapted to the study gratings composed of periodic infinite cylinders. However, there are fundamental differences in the mathematics, since this problem is usually addressed by solving 2-dimensional Helmholtz equations. We therefore neglect this problem in order to concentrate on the fully 3-dimensional problem of particulate gratings like those of figure 6.1.

The first five sections constitute the heart of this chapter since they describe the general mathematical analysis of gratings using spherical harmonic lattice sums. Sections 6.6 and 6.7 treat numerical methods for calculating lattice sums, while addition theorems and numerical methods for special functions are treated in sections 6.8 and 6.9 respectively. Support material, erratum, and recent advances will be made available on my website: www.fresnel.fr/perso/stout/index.htm.

6.2 Waves and partial waves

A fundamental aspect of the particulate gratings is that they can be viewed as a multiple-scattering phenomenon with light propagating through the host medium between individual scatterings events. The wave equation for light in this homogeneous isotropic medium is:

$$\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{E}+k^2\boldsymbol{E}=\boldsymbol{0}\,, \tag{6.1}$$

where $k=\sqrt{\varepsilon_b\mu_b}\frac{\omega}{c}$ is the wavenumber of the host or "background" medium. Solutions of eq.(6.1) satisfy both the vector Helmholtz equation,

$$\Delta\boldsymbol{E}+k^2\boldsymbol{E}=\boldsymbol{0}\,, \tag{6.2}$$

and the additional constraint that the longitudinal field components are null:

$$\boldsymbol{\nabla}\cdot\boldsymbol{E}=0\,. \tag{6.3}$$

A basis set for solutions to the vector Helmholtz equation of eq.(6.2) can be readily constructed starting from the scalar Helmholtz equation:

$$\Delta\Phi+k^2\Phi=0\,. \tag{6.4}$$

As well established in textbooks[10], eq.(6.4) can be solved by separation of variables in spherical coordinates with 'regular' solutions taking the form of spherical harmonics, $Y_{n,m}$ multiplied by spherical Bessel functions, $j_n(kr)$, that are finite valued for all values of r. There also exists linearly indpendent 'irregular' solutions to this equation, called spherical Neumann functions, $y_n(kr)$, which possess essential singularities for $kr \to 0$. Details concerning the properties and calculation of the $Y_{n,m}(\theta,\phi)$ are given in section 6.8.1.

The spherical coordinate solutions to eq.(6.4) will henceforth be referred to as Cartesian partial waves and will be defined as:

$$\mathscr{J}_{n,m}(k\boldsymbol{r}) \equiv j_n(kr)\,Y_{n,m}(\widehat{\boldsymbol{r}}) \ , \quad \text{and} \quad \mathscr{Y}_{n,m}(k\boldsymbol{r}) \equiv y_n(kr)\,Y_{n,m}(\widehat{\boldsymbol{r}}) \ . \tag{6.5}$$

The regular partial waves, $\mathscr{J}_{n,m}$, can serve as a basis set for any source-free incident field solution to eq.(6.4). Outgoing partial waves solutions of the Helmholtz equation, denoted $\mathscr{H}_{n,m}$, will be of primary interest in grating theory since they will be used to describe fields scattered by the grating. They are defined as a superposition of the regular and irregular partial waves:

$$\mathscr{H}_{n,m}(k\boldsymbol{r}) \equiv h_n(kr)\,Y_{n,m}(\widehat{\boldsymbol{r}}) = \mathscr{J}_{n,m}(k\boldsymbol{r}) + i\mathscr{Y}_{n,m}(k\boldsymbol{r}) \ . \tag{6.6}$$

Incident field solutions to the vector Helmholtz equation of eq.(6.2) can be expressed as Cartesian partial waves associated with unit vectors along each axis i.e.:

$$\boldsymbol{E}_{inc}(\boldsymbol{r}) = \widehat{\boldsymbol{x}}\sum_{n,m}\alpha^{(x)}_{n,m}\mathscr{J}_{n,m}(k\boldsymbol{r}) + \widehat{\boldsymbol{y}}\sum_{n',m'}\alpha^{(y)}_{n',m'}\mathscr{J}_{n',m'}(k\boldsymbol{r}) + \widehat{\boldsymbol{z}}\sum_{n'',m''}\alpha^{(z)}_{n'',m''}\mathscr{J}_{n'',m''}(k\boldsymbol{r}) \ . \tag{6.7}$$

The field scattered field scattered by a particle in the context of the vector Helmholtz equation can likewise be developed in terms of the outgoing spherical waves:

$$\boldsymbol{E}_{scat}(\boldsymbol{r}) = \widehat{\boldsymbol{x}}\sum_{n,m}\beta^{(x)}_{n,m}\mathscr{H}_{n,m}(k\boldsymbol{r}) + \widehat{\boldsymbol{y}}\sum_{n',m'}\beta^{(y)}_{n',m'}\mathscr{H}_{n',m'}(k\boldsymbol{r}) + \widehat{\boldsymbol{z}}\sum_{n'',m''}\beta^{(z)}_{n'',m''}\mathscr{H}_{n'',m''}(k\boldsymbol{r}) \ , \tag{6.8}$$

provided that the origin, $\boldsymbol{r} = \boldsymbol{0}$, is chosen to lie inside the particle and that the field description is applied only to regions lying outside the particle. The field in eq.(6.8) represents the field scattered by a single scatterer, so the grating problem in terms of partial waves must sum over the field scattered by all the particles in the lattice. Finding efficient ways for calculating the lattice sum will therefore figure prominently in the subsequent sections of this chapter.

Before studying T-matrices in the next section, we first address an important technical issue. The field expansions in eq.(6.7) and eq.(6.8) have both transverse and longitudinal components and therefore are not generally solutions of the light propagation problem of eq.(6.1). The transverse wave condition of eq.(6.3) can be satisfied by requiring that the Cartesian field coefficients, $\alpha^{(x,y,z)}_{n,m}$, (and respectively $\beta^{(x,y,z)}_{n,m}$) satisfy certain relations amongst themselves. The important point is to remark that the constraint conditions, although somewhat complex in spherical coordinates, only affect the partial wave *coefficients*, and not the partial waves themselves. Consequently, in the rest of this chapter, we can generally restrain our attention to lattice sums of scalar partial wave sums even though the end goal is to describe electromagnetic field scattering.

Expressing the transverse vector partial waves in terms of the Cartesian partial waves of eq.(6.7) or eq.(6.8) is a relatively complex but straight forward affair involving angular momentum coupling formalism, coordinate transformations, and recurrence relations.[21] Consequently, it is more common to invoke one of the various methods that have been devised over

the years to directly generate transverse partial waves: Debye potentials, Hertz potentials, the Boulenkamp-Casimir approach[10], pilot vector techniques[5], etc. Whatever one's "preferred" technique and notation, the two types of transverse partial waves, $\mathbf{\Psi}_{\mathscr{J},q,p}$ (often denoted $\boldsymbol{M}_{\mathscr{J},p}$ and $\boldsymbol{N}_{\mathscr{J},p}$ in the literature), can be expressed:

$$\begin{aligned}\mathbf{\Psi}_{\mathscr{J},1,p}(k\boldsymbol{r}) &\equiv j_n(kr)\boldsymbol{X}_{n,m}(\widehat{\boldsymbol{r}}) \\ \mathbf{\Psi}_{\mathscr{J},2,p}(k\boldsymbol{r}) &\equiv \frac{1}{kr}\left\{\sqrt{n(n+1)}\,j_n(kr)\boldsymbol{Y}_{n,m}(\widehat{\boldsymbol{r}}) + [krj_n(kr)]'\boldsymbol{Z}_{n,m}(\widehat{\boldsymbol{r}})\right\}\,,\end{aligned} \tag{6.9}$$

where $\boldsymbol{X}_{n,m}$, $\boldsymbol{Y}_{n,m}$, and $\boldsymbol{Z}_{n,m}$ are the *vector* spherical harmonics (VSHs), (described in section eq.(6.9.3)). The first subscript, $\mathscr{J}$, on $\mathbf{\Psi}_{\mathscr{J},q,p}$ serves to indicate that the radial dependence is governed by spherical Bessel functions. A value of $q=1$ in the second subscript indicates a transverse electric (TE) wave (*i.e.* possessing no radial electric field component), while $q=2$ indicates a transverse magnetic (TM) wave having no radial magnetic field. In order to minimize the number of subscripts, we adopt the common procedure that the third subscript p of $\mathbf{\Psi}_{\mathscr{J},q,p}$, replaces the *two* multipole subscripts n and m by defining its value such that[31]:

$$p(n,m) \equiv n(n+1) - m\,. \tag{6.10}$$

With the notation of eq.(6.9), one can express any incident field satisfying equation (6.1) in terms of the transverse vector partial waves:

$$\boldsymbol{E}_{\text{inc}}(\boldsymbol{r}) = E\sum_{q=1,2}\sum_{p=1}^{\infty}\mathbf{\Psi}_{\mathscr{J},q,p}(k\boldsymbol{r})\,a_{q,p}\,, \tag{6.11}$$

where $a_{q,p}$ are (dimensionless) field coefficients, and E is a constant with the dimension of electric field and which can be used to adjust the field strength. With this notation, the field scattered by a particle whose circumscribing sphere is centered at a position $\boldsymbol{x}_j$ can be written:

$$\boldsymbol{E}_{\text{scat}}(\boldsymbol{r}_j) = E\sum_{q=1,2}\sum_{p=1}^{\infty}\mathbf{\Psi}_{\mathscr{H},q,p}(k\boldsymbol{r}_j)\,f_{q,p}^{(j)}\,, \tag{6.12}$$

where $\boldsymbol{r}_j \equiv \boldsymbol{r} - \boldsymbol{x}_j$, and $f_{q,p}^{(j)}$ are the scattering coefficients of the particle j. The index, $\mathscr{H}$, on the $\mathbf{\Psi}_{\mathscr{H},q,p}$ indicates that the radial dependence should be governed by spherical Hankel functions, $h_n(kr)$, rather than the spherical Bessel functions, $j_n(kr)$, found in the $\mathbf{\Psi}_{\mathscr{J},q,p}$ functions of eq.(6.9).

6.3 T-matrix theory

6.3.1 Green functions and T-matrices

The fundamental object that one would like to calculate in a multiple-scattering system (like a particulate grating) is the system Green's function. However, the dyadic Green's function for a homogeneous medium has a strongly singular behavior and needs to be defined in the context of distributions.[5] The T-matrix formalism allows us to largely circumvent this singular behavior, and also to work directly in terms of fields which is often more manageable than the

relatively intricate dyadic Green's function formalism. For instance, the operator form of the Green function of a single object in a homogeneous medium can be written:

$$\boldsymbol{G} = \boldsymbol{g} + \boldsymbol{g}\boldsymbol{t}\boldsymbol{g} \,, \tag{6.13}$$

where $\boldsymbol{t}$ is the isolated particle (or *1-body*) T-matrix operator, and $\boldsymbol{g}$ is the Green function operator of the homogeneous medium (sometimes called a *propagator*). In this formalism, the singular behavior is relegated to the propagator, $\boldsymbol{g}$, leaving the (non-singular) scattering response due to the object being described by $\boldsymbol{t}$.

Furthermore, when considering excitations outside the scatterer, the homogeneous Green function, $\boldsymbol{g}$ to the right of the $\boldsymbol{t}$ operator acting on the sources generates the incident field, while the $\boldsymbol{g}$ to the left of it generates the scattered field.[24] In the partial wave basis, $\boldsymbol{t}$ then truly takes the form of a *matrix*, henceforth denoted, t, that relates the incident field coefficients to the scattered field coefficients:

$$f = ta \,, \tag{6.14}$$

where a and f are column matrices composed respectively of the incident field and scattered field coefficients (cf. eqs.(6.11 and (6.12)).[30] The 1-body T-matrix, t, in this expression is now truly a *matrix* relating field coefficients of partial wave field decompositions.

This T-matrix formalism can be extended to include systems containing N particles. The system Green function can be written,

$$\boldsymbol{G} = \boldsymbol{g} + \boldsymbol{g}\left(\sum_{j=1}^{N} \boldsymbol{T}^{(j)}\right)\boldsymbol{g} \,, \tag{6.15}$$

where the *multiple-scattering* (or N-body) T-matrix operators, $\boldsymbol{T}^{(j)}$, are associated with each particle and which incorporates all the multiple-scattering effects due to the presence of the $N-1$ other particles in the system. Passing once again to a partial wave field description, the multiple-scattering T-matrix, $T^{(j)}$, generates the field scattered by each particle in terms of the field incident on the system:

$$f^{(j)} = T^{(j)} a^{(j)} \,, \tag{6.16}$$

where $a^{(j)}$ indicates the incident field developed on a coordinate system centered on the j^{th} particle. All multiple-scattering phenomenon and some rather subtle technical difficulties have all been incorporated into the definition of $T^{(j)}$, but nowadays they can be calculated rather readily for systems with a finite number of particles starting from the 1-body T-matrices, $t^{(j)}$, of the individual particles.[30]

The number of particles in a grating problem is infinite (from the ideal mathematical standpoint), which given their physical content would render exact calculations of $T^{(j)}$ impossible. Nevertheless, the fact that the system is identical when viewed from any lattice site allows the $T^{(j)}$ matrices to be calculated as a lattice sum as we shall see in section 6.3.4. We first rapidly review below our notation and terminology for lattices.

6.3.2 Mie theory T-matrices

Since the T-matrix of individual scatterers is not the goal of this chapter, we will principally consider examples in which the particles in the grating of chain are isotropic homogeneous

spheres with relative constitutive parameters, ε_s and/or μ_s, that differ from that of the external "background" medium, ε_b and μ_b. The principle advantage of this choice is that the matrix elements of the T-matrix are determined analytically from Mie theory. This of course permits precise calculations, but it also allows for an understanding of the relationship between multipole order and particle size which is important in numerical calculations as reviewed in section 6.3.5, and discussed in detail in ref[30].

For spherically symmetric scatterers, the scattering coefficients are directly proportional to the excitation coefficients of the same order which reads

$$f_{q,p} = t_{q,n} e_{q,p} \ , \tag{6.17}$$

where the coefficients $t_{q,n}$ depend only on the orbital quantum number, n, and on the field character ($q = 1$ for TE waves and $q = 2$ for TM waves). A comparison of eq.(6.17) with eq.(6.14) shows that the individual T-matrix for a spherically symmetric scatterer is a diagonal matrix,

$$[t]_{q,p;q',p'} = \delta_{q,q'} \delta_{p,p'} t_{q,n} \ . \tag{6.18}$$

The values of $t_{q,n}$ are determined by developing the field inside and outside the homogeneous sphere in terms of the transverse vector partial waves (cf. eq.(6.9)) while imposing the continuity of the tangential electric and magnetic fields at the surface of the sphere. Although the Mie T-matrix elements are most frequently expressed in the form originally given by Mie (cf. Bohren and Huffman[**?**] and eq.(6.20 below), these expressions tend to hide the duality symmetry between the TE (magnetic) and TM (electric) matrix elements. Consequently, we prefer the following expressions:

$$\begin{aligned} t_{1,n} &= \frac{j_n(kR)}{h_n(kR)} \left\{ \frac{\frac{\mu_s}{\mu_b}\varphi_n(kR) - \varphi_n(k_sR)}{\varphi_n(k_sR) - \frac{\mu_s}{\mu_b}\varphi_n^{(3)}(kR)} \right\} \qquad \text{(TE)} \\ t_{2,n} &= \frac{j_n(kR)}{h_n(kR)} \left\{ \frac{\frac{\varepsilon_s}{\varepsilon_b}\varphi_n(kR) - \varphi_n(k_sR)}{\varphi_n(k_sR) - \frac{\varepsilon_s}{\varepsilon_b}\varphi_n^{(3)}(kR)} \right\} \qquad \text{(TM)} \ , \end{aligned} \tag{6.19}$$

which have more transparently symmetric expressions with respect to their respectively TE and TM natures of the coefficients. There expressions also have a numerical advantage since the $\varphi_n(z)$ and $\varphi_n^{(3)}(z)$ and $j_n(z)/h_n(z)$ functions can all be rapidly evaluated via simple recursion relations and have well behaved limit behaviors as shown in eqs.(6.180), (6.176) (6.180) and (6.192) of section 6.9.2.

Traditionally, the Mie coefficients are denoted a_n and b_n and are of opposite in sign from the T-matrix elements, *i.e.*:

$$\begin{aligned} a_n &= \frac{\frac{\mu_s}{\mu_b}\psi_n(kR)\,\psi_n'(k_sR) - \rho\,\psi_n(k_sR)\,\psi_n'(kR)}{\frac{\mu_s}{\mu_b}\xi_n(kR)\,\psi_n'(k_sR) - \rho\,\psi_n(k_sR)\,\xi_n'(kR)} = -t_{2,n} \\ b_n &= \frac{\frac{\mu_s}{\mu_b}\psi_n(k_sR)\,\psi_n'(kR) - \rho\,\psi_n(kR)\,\psi_n'(k_sR)}{\frac{\mu_s}{\mu_b}\psi_n(k_sR)\,\xi_n'(kR) - \rho\,\xi_n(kR)\,\psi_n'(k_sR)} = -t_{1,n} \ , \end{aligned} \tag{6.20}$$

where $\rho \equiv k_s/k = n_s/n$, denotes the refractive index contrast between the sphere and the background media, and $\psi_n(z) \equiv z j_n(z)$ and $\xi_n(z) \equiv z h_n(z)$ are respectively the Ricatti forms of the

spherical Bessel and Hankel functions. These more widely used expressions for the Mie coefficients are have a somewhat more symmetric appearance in optics where there is usually no permeability contrast, *i.e.* $\mu_s/\mu_b = 1$ as was assumed by Mie and most other authors in optics. Nevertheless, our experience is that the expressions in eq.(6.19) have a more transparent physical interpretations and numerical properties.

6.3.3 Direct and reciprocal lattices

A lattice, Λ, of dimension d_Λ, is invariant under a coordinate system translation along any vector, $\boldsymbol{r_j}$, that can be expressed

$$\boldsymbol{r_j} = \sum_{i=1}^{d_\Lambda} j_i \boldsymbol{a}_i \ , \tag{6.21}$$

where $\boldsymbol{a}_i$ are the *primitive lattice vectors*, and $\boldsymbol{j} \equiv (j_1, ..., j_{d_\Lambda})$ is a shorthand notation for a set of d_Λ relative integers, $j_i \in \mathbb{Z}$. In order to diminish the number of subscripts, we will sometimes employ an alternative notation for the primitive lattice vectors: $\boldsymbol{a} \equiv \boldsymbol{a}_1$, $\boldsymbol{b} \equiv \boldsymbol{a}_2$, and $\boldsymbol{c} \equiv \boldsymbol{a}_3$. It also proves convenient to define the x and y axis of the system so that the primitive lattice vectors can be expressed: $\boldsymbol{a} = (a, 0, 0)$, $\boldsymbol{b} = (b_x, b_y, 0)$, and $\boldsymbol{c} = (c_x, c_y, c_z)$.

When $d_\Lambda = 3$, the $\boldsymbol{r_j}$ ensemble defines a crystalline type lattice, henceforth denoted (L), as frequently encountered in photonic crystals and "meta-materials". A two-dimensional grating, or mono-layer lattice (ML), like that of figure 6.1, occurs when the system invariance only occurs for 2D displacements of $\boldsymbol{r_j} = j_a\boldsymbol{a} + j_b\boldsymbol{b}$. Finally, linear chains (C) are only invariant with respect to translations of $\boldsymbol{r}_j = j\boldsymbol{a}$.

The *reciprocal* lattice, Λ^*, is defined in terms of lattice 'wave-vectors', $\boldsymbol{p_g}$, defined in terms of the primitive reciprocal lattice vectors, $\widetilde{\boldsymbol{a}}_i$:

$$\boldsymbol{p_g} = 2\pi \sum_{i=1}^{d_\Lambda} g_i \widetilde{\boldsymbol{a}}_i \ , \tag{6.22}$$

where $g_i \in \mathbb{Z}$. The primitive reciprocal vectors, $\widetilde{\boldsymbol{a}}_j$, are defined such that their scalar products with respect to $\boldsymbol{a}_j$ satisfy:

$$\boldsymbol{a}_i \cdot \widetilde{\boldsymbol{a}}_j = \delta_{ij} \qquad i, j = 1, ..., d_\Lambda \ . \tag{6.23}$$

From eqs.(6.21) (6.22) and (6.23), one readily finds that the reciprocal lattice vectors, $\boldsymbol{p_g}$, of eq.(6.22), have the property

$$\boldsymbol{r_j} \cdot \boldsymbol{p_g} = 2\pi N \ , \tag{6.24}$$

where N is some integer (which results in $\exp\left(i\boldsymbol{r_j} \cdot \boldsymbol{p_g}\right) = 1$ for all $\boldsymbol{r_j}$ and $\boldsymbol{p_g}$).

The unit cell "volume", $\mathscr{A}$, appears repeatedly in theories of particulate lattices. For lattice dimensions of $d_\Lambda = 1$, 2, and 3, the corresponding $\mathscr{A}_{1,2,3}$ is given by:

$$\begin{cases} \mathscr{A}_1 = |\boldsymbol{a}| & d_\Lambda = 1 \\ \mathscr{A}_2 = |\boldsymbol{a} \times \boldsymbol{b}| & d_\Lambda = 2 \\ \mathscr{A}_3 = |(\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{c}| & d_\Lambda = 3 \end{cases} \ , \tag{6.25}$$

with dimensions of "length" for $d_\Lambda = 1$, "area" for $d_\Lambda = 2$ and "volume" for $d_\Lambda = 3$. The corresponding "volume" of the reciprocal space lattice sites are given by $\mathscr{A}^{-1}$.

6.3.4 Grating T-matrices

Each site of a lattice is identical to all the others so that the multiple scattering T-matrices of eq.(6.16) are all equal, i.e. $T^{(j)} = T$. The scattering coefficients $f^{(j)}$ are then given by:

$$f^{(j)} = Ta^{(j)} \ . \tag{6.26}$$

The trouble with this equation is that the coefficients $f^{(j)}$ and $a^{(j)}$ are expressed on a localized partial wave basis, but the long range nature of scattered fields would require the T-matrices to act on very high multipole orders in order to account for these long-rang interactions.

Since manipulating high multipole orders is numerically inefficient, one considerably simplifies this problem by only calculating the multiple-scattering T-matrices for incident fields satisfying a quasi-periodic condition. Quasi-periodicity can be viewed as requiring the partial-wave decomposition of the incident field on each lattice site, $\boldsymbol{r_j}$, to satisfy,

$$a^{(j)} = e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} a \ , \tag{6.27}$$

where 'a' corresponds to the incident field coefficients at the origin, and $\boldsymbol{\beta}$, the on-shell quasi-periodicity vector. The term 'on-shell' indicates that the quasi-periodic vector satisfy $\boldsymbol{\beta}^2 = k^2$ since the incident field must satisfy Eq.(6.1) in the external medium.

The quasi-periodic condition can be viewed as a partial Fourier transform description in that the overall field behavior is of an oscillatory nature, while the quasi-periodic T-matrix, $T_{\boldsymbol{\beta}}$, describes local-field perturbations due to the presence of the particles. Consequently, one expects the $T_{\boldsymbol{\beta}}$ matrices to be well approximated on a truncated (*i.e.* finite) partial-wave basis (similar to the behavior of the isolated particle T-matrices[30]). The quasi-periodic condition allows the Foldy-Lax equations for the multiple-scattering T-matrices to take the form:

$$T_{\boldsymbol{\beta}} = t + t\Omega_{\boldsymbol{\beta}} T_{\boldsymbol{\beta}} \ , \tag{6.28}$$

where the $\Omega_{\boldsymbol{\beta}}$ matrix designates a quasi-periodic lattice sum of the irregular translation-addition matrices:

$$\Omega_{\boldsymbol{\beta}}(k) = \sum_{\substack{\boldsymbol{r_j}\in\Lambda \\ \boldsymbol{r_j}\neq\boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} H\left(k\boldsymbol{r_j}\right) \ . \tag{6.29}$$

The analytical properties of the irregular translation-addition matrix, $H(\boldsymbol{x})$, are described in section 6.8.3 where one also gives expressions for its matrix elements.

The exclusion of the 'origin' lattice site, $\boldsymbol{r_j} = \boldsymbol{0}$, from the sum in $\Omega_{\boldsymbol{\beta}}$ has a physical significance in that it accounts for propagation of the light scattered by all the other particles in the lattice onto the particle at the origin (the light 'scattered' by the particle onto itself has already been included in the individual T-matrix, t). One finds in section 6.8.3 that each matrix element of the translation-addition matrix, $H\left(k\boldsymbol{r_j}\right)$, can be written:

$$\left[H\left(k\boldsymbol{r_j}\right)\right]_{p,q;p',q'} = \sum_{l,m} C_{l,m}\left(p,q;p',q'\right) h_l\left(kr_j\right) Y_{l,m}\left(\widehat{\boldsymbol{r}}_{\boldsymbol{j}}\right) \ , \tag{6.30}$$

where the sum over the multipole indices, (l,m) is finite. Expressions for the $C_{l,m}(p,q;p',q')$ coefficients[31, 5, 29] are given in the section 6.8. Inserting eq.(6.30) into eq.(6.29) and rear-

ranging the summations, we find

$$\left[\Omega_{\boldsymbol{\beta}}\right]_{p,q;p',q'} = \sum_{l,m} C_{l,m}\left(p,q;p',q'\right) \sum_{\substack{\boldsymbol{r_j}\in\Lambda \\ \boldsymbol{r_j}\neq\boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} h_l\left(kr_j\right) Y_{l,m}\left(\widehat{\boldsymbol{r}}_{\boldsymbol{j}}\right)$$
$$\equiv \sum_{l,m} C_{l,m}\left(p,q;p',q'\right) S_{l,m}\left(k,\boldsymbol{\beta}\right) \ . \tag{6.31}$$

where we have defined $S_{l,m}\left(k,\boldsymbol{\beta}\right)$ as a Hankel function lattice sum such that:

$$S_{n,m}\left(k,\boldsymbol{\beta}\right) \equiv S^{\mathscr{H}}_{n,m}\left(k,\boldsymbol{\beta}\right) \equiv \sum_{\substack{\boldsymbol{r_j}\in\Lambda \\ \boldsymbol{r_j}\neq\boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} \mathscr{H}_{n,m}\left(k\boldsymbol{r_j}\right) \ . \tag{6.32}$$

We recall that $\mathscr{H}_{n,m}\left(\boldsymbol{x}\right)$ was defined in eq.(6.6) as a partial wave of the spherical Hankel function type.

It will sometimes prove useful to calculate the analogous lattice sums over the partial waves of the Bessel or Neumann types, denoted respectively, $S^{\mathscr{J}}_{n,m}\left(k,\boldsymbol{\beta}\right)$ and $S^{\mathscr{Y}}_{n,m}\left(k,\boldsymbol{\beta}\right)$. Since we will principally be concerned with the partial wave lattice sums of the Hankel function type, $S_{n,m}$ without a superscript will always indicate a lattice sum of the Hankel function type. We also remark that the exclusion of the origin position from the lattice sum is important from a mathematical standpoint since the Hankel functions have an essential singularity at their origin.

The solution to eq.(6.28) for the multiple-scattering T-matrix is readily formulated in terms of matrix inversion:

$$T_{\boldsymbol{\beta}} = \left[t^{-1} - \Omega_{\boldsymbol{\beta}}\right]^{-1} \ . \tag{6.33}$$

Once the $T_{\boldsymbol{\beta}}$ matrix is known, the scattering field coefficients for any particle, j, in the lattice is the same as the coefficients at the origin but multiplied by a $e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}}$ phase factor. In the matrix notation, this is simply expressed:

$$f^{(j)}_{\boldsymbol{\beta}} = e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} f_{\boldsymbol{\beta}} = e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} T_{\boldsymbol{\beta}} a \ , \tag{6.34}$$

where a is the column matrix composed of the incident field coefficients developed around the origin.

6.3.4.1 Far-fields

The field 'scattered' by the grating (i.e. 'transmitted' and 'reflected' diffraction orders) can be determined by inserting eq.(6.34) into eq.(6.12) wherein the scattered field takes the form of a lattice sum of the transverse-outgoing-*vector* partial waves, $\boldsymbol{\Psi}_{\mathscr{H},q,p}$, described in eq.(6.9) of section 6.2:

$$\boldsymbol{E}_{\mathrm{s},\Lambda}\left(\boldsymbol{r}\right) = E \sum_{\boldsymbol{r_j}\in\Lambda} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}} \boldsymbol{\Psi}_{\mathscr{H}}\left(k\boldsymbol{r_j}\right) f_{\boldsymbol{\beta}}$$
$$\equiv E \sum_{q=1,2} \sum_{p=1}^{\infty} \left[\sum_{\boldsymbol{r_j}\in\Lambda} \boldsymbol{\Psi}_{\mathscr{H},q,p}\left(k\boldsymbol{r_j}\right) e^{i\boldsymbol{\beta}\cdot\boldsymbol{r_j}}\right] \left[f_{\boldsymbol{\beta}}\right]_{q,p} \ . \tag{6.35}$$

We will see in eq.(6.66) of section 6.4.4 that for each multipole order, $p = 1,...\infty$, and transverse field character, $q = 1,2$; the term in brackets can be re-expressed as an infinite sum of plane waves. Only a finite subset of these waves are of the propagative type however (the rest are all of an evanescent nature). Consequently, the multipole summation of eq.(6.35) allows one to calculate the efficiency of each reflected or transmitted order in the far field.

6.3.4.2 Near-fields

Another quantity of physical interest is that of near fields in a particulate grating (non-linear effects, SERS, etc.). The plane wave expansion discussed above for far fields could be invoked in principal, but for near fields one must also calculate the (infinite) evanescent orders that one could neglect in the far field. The convergence of the plane wave expansion will generally be poor near the grating, which renders this approach unattractive.

As long as the incident field is quasi-periodic with respect to the grating, one needs only to determine the near fields in a single Brillouin zone around a given lattice site (the site at the origin being the most practical). In this case, it seems clear that the localized multipolar field developments are well adapted to the development of the local field in the Brillouin zone. In multiple scattering theory, the $f_{\boldsymbol{\beta}}$ coefficients give the field scattered by the particle at the origin, while the *excitation* field corresponds to the field at that was 'incident' on this particle, *i.e.* the superposition of the field incident on the grating and the field scattered by all the other particles in the system. This excitation field can be developed on the regular partial waves and its coefficients, $e_{\boldsymbol{\beta}}$, related to the scattering coefficients via the 1-body T-matrix via the relation:

$$e_{\boldsymbol{\beta}} = t^{-1} f_{\boldsymbol{\beta}} \ . \tag{6.36}$$

The total field in the Brillouin zone is simply a superposition of the scattered and excitation field:

$$\begin{aligned} \boldsymbol{E}_{\mathrm{t}}^{(\mathrm{B.z.})}(\boldsymbol{r}) &= E\left(\boldsymbol{\Psi}_{\mathscr{H}}(k\boldsymbol{r})\, f_{\boldsymbol{\beta}} + \boldsymbol{\Psi}_{\mathscr{J}}(k\boldsymbol{r})\, t^{-1} f_{\boldsymbol{\beta}}\right) \\ &\equiv E \sum_{q=1,2} \sum_{p=1}^{\infty} \left[\boldsymbol{\Psi}_{\mathscr{H},q,p}(k\boldsymbol{r})\, f_{q,p} + \boldsymbol{\Psi}_{\mathscr{J},q,p}(k\boldsymbol{r})\, e_{q,p}\right] \ . \end{aligned} \tag{6.37}$$

6.3.4.3 Propagating modes

When the quasi-periodic incident field nearly matches a true guided mode of a structure and/or quasi-modes (also referred to as *leaky* modes), then the response of the structure will tend to be dominated by these 'modes'. True propagating modes can be guided by either 1, 2 or 3-D lattices provided that the particles are free from intrinsic losses, however true propagating modes in 1D and 2D periodic will require the quasi-periodic vector to have evanescent behavior in the dimensions perpendicular to the lattice. Lossless 3D lattices on the other hand can have modes described by entirely real values of $\boldsymbol{\beta}$. In the presence of intrinsic losses however, all propagating modes will be a leaky nature since energy is lost during propagation. Such leaky modes can be described by a complex valued $\boldsymbol{\beta}$-vector or a complex value of frequency. The determination of a 'leaky' mode thus involves searching for a complex pole in the determinant of the multiple-scattering T-matrix, $\left|T_{\boldsymbol{\beta}}\right|$.

Since matrix inversions are numerically expensive, one may prefer to look for zero eigenvectors, ν_α, of the matrix $\left[t^{-1} - \Omega_{\boldsymbol{\beta}_\alpha}\right]$, *i.e.*

$$\left[t^{-1} - \Omega_{\boldsymbol{\beta}_\alpha}\right] \nu_\alpha = 0 \; . \tag{6.38}$$

However, the search for zero eigenvalues can limit the implantation of complex analysis methods that have proven useful in determining the position of poles in the complex plane.

The Floquet mode associated with the eigenvector, ν_α, can be constructed from eq.(6.38) coupled with the plane-wave development the terms in eigenvector, ν_α:

$$\boldsymbol{E}_\alpha^{(\mathrm{F.m.})}(\boldsymbol{r}) = E \sum_{q=1,2} \sum_{p=1}^{\infty} \left[\sum_{\boldsymbol{r_j} \in \Lambda} \boldsymbol{\Psi}_{\mathscr{H},q,p}\left(k\boldsymbol{r_j}\right) e^{i\boldsymbol{\beta}_\alpha \cdot \boldsymbol{r_j}} \right] [\nu_\alpha]_{q,p} \; . \tag{6.39}$$

Before finishing this section, it should be pointed out that matrix inversion solutions to the multiple scattering problem (like that given in eq.(6.33)) were disregarded for a long time in favor of iterative solutions to the T-matrix or underlying linear system of equations. The reason for this is that the matrix $\left[t^{-1} - \Omega_{\boldsymbol{\beta}_\alpha}\right]$ is generally ill-conditioned. This difficulty can be generally overcome by analytical matrix balancing as described in the next section 6.3.5.

6.3.5 Matrix balancing

Although not necessary from a formal standpoint, analytical matrix balancing improves the conditioning of the matrices occurring in multiple-scattering calculations for both matrix inversion and eigenvalue resolution.[30] Analytical matrix balancing can be achieved by multiplying a matrix from both the right and left by diagonal matrices, $[\xi]$ and $[\psi]^{-1}$, whose matrix elements are given by:

$$[\psi]_{q,q',p,p'} = \delta_{q,q'}\delta_{p,p'}\psi_n(kR) \; , \qquad [\xi]_{q,q',p,p'} = \delta_{q,q'}\delta_{p,p'}\xi_n(kR) \; , \tag{6.40}$$

where $\psi_n(kR)$ and $\xi_n(kR)$ are respectively the regular and irregular spherical Ricatti-Bessel functions (cf. 6.172) and R the radius of the minimal circumscribing sphere surrounding the scatterers.

Matrix balancing can be readily formulated by defining normalized incident and scattering coefficients, $\overline{a}$ and $\overline{f}$ respectively such that:

$$\overline{a} \equiv [\psi]\, a \; , \qquad \overline{f}_{\boldsymbol{\beta}} \equiv [\xi]\, f_{\boldsymbol{\beta}} \; . \tag{6.41}$$

The associated normalized or 'balanced' matrices are defined[30]:

$$\overline{t} \equiv [\xi]\, t\, [\psi]^{-1} \; , \qquad \overline{T}_{\boldsymbol{\beta}} \equiv [\xi]\, T_{\boldsymbol{\beta}}\, [\psi]^{-1} \; , \qquad \overline{\Omega}_{\boldsymbol{\beta}} \equiv [\psi]\, \Omega_{\boldsymbol{\beta}}\, [\xi]^{-1} \; . \tag{6.42}$$

The above definitions were chosen such that eqs.(6.26) and (6.28) respectively take the form:

$$\overline{f}^{(j)} \equiv \overline{T}_{\boldsymbol{\beta}} \overline{a}^{(j)} \; , \tag{6.43}$$

and

$$\overline{T}_{\boldsymbol{\beta}} \overline{a} = \overline{t}\overline{a} + \overline{t}\overline{\Omega}_{\boldsymbol{\beta}} \overline{T}_{\boldsymbol{\beta}} \overline{a} \; . \tag{6.44}$$

The normalized T-matrix, $\overline{T}$, is then obtained via the generally well-conditioned matrix inversion:

$$\overline{T}_{\boldsymbol{\beta}} = \left[\overline{t}^{-1} - \overline{\Omega}_{\boldsymbol{\beta}}\right]^{-1} . \tag{6.45}$$

Since we generally want the non-normalized T-matrix for applications, we reconstruct, $T_{\boldsymbol{\beta}}$ via a final multiplication by our diagonal matrices:

$$T_{\boldsymbol{\beta}} = [\xi]^{-1}\, \overline{T}_{\boldsymbol{\beta}}\, [\psi] . \tag{6.46}$$

6.4 Mathematical relations for lattice sums

This section is dedicated to reviewing the mathematical relations that allow one to treat lattice sums for lattices of dimensions $d_\Lambda = 1,2,3$ (i.e. particulate chains, gratings, and crystals). They will notably allow us to evaluate the lattice sum in eq.(6.31) which is used to calculate far-field response from gratings. These relations were derived (and often rederived) in many places, and we refer the reader to refs.[15, 16, 19, 20, 8] for additional details and perspectives.

The most difficult mathematical problem to address will be the evaluation of Hankel function lattice sum, $S^{\mathscr{H}}_{n,m}$ that was introduced in eq.(6.32) for the calculation of the $\Omega_{\boldsymbol{\beta}}$ matrix of eq.(6.31).

$$S^{\mathscr{H}}_{n,m} \equiv \sum_{\substack{\boldsymbol{r}_j \in \Lambda \\ \boldsymbol{r}_j \neq \boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}_j}\, \mathscr{H}_{n,m}\left(k\boldsymbol{r}_j\right) \tag{6.47}$$

When not otherwise specified, partial waves lattice sums will always be assumed to be of the Hankel function type. The underlying reason for this appears in the translation-addition where Hankel functions allow one to re-express waves *scattered* by a given lattice site as waves *incident* on a different lattice site.

We will occasionally consider Bessel and Neumann types of scalar partial waves:

$$\begin{aligned} S^{\mathscr{Y}}_{n,m} &\equiv \sum_{\substack{\boldsymbol{r}_j \in \Lambda \\ \boldsymbol{r}_j \neq \boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}_j}\, \mathscr{Y}_{n,m}\left(k\boldsymbol{r}_j\right) \\ S^{\mathscr{J}}_{n,m} &\equiv \sum_{\substack{\boldsymbol{r}_j \in \Lambda \\ \boldsymbol{r}_j \neq \boldsymbol{0}}} e^{i\boldsymbol{\beta}\cdot\boldsymbol{r}_j}\, \mathscr{J}_{n,m}\left(k\boldsymbol{r}_j\right) , \end{aligned} \tag{6.48}$$

The interest of these sums in part is due to the fact that $S^{\mathscr{H}}_{n,m} = S^{\mathscr{J}}_{n,m} + iS^{\mathscr{Y}}_{n,m}$ but also because $S^{\mathscr{J}}_{n,m}$ can potentially prove useful in certain applications. We will see that the relations developed in this chapter permit the $S^{\mathscr{J}}_{n,m}$ sum to be evaluated in closed form, but unfortunately the Neumann partial wave sum, $S^{\mathscr{Y}}_{n,m}$, appears to be as difficult to evaluate as the Hankel partial wave sum, and a closed form expression does not appear to be possible.

6.4.1 Lattice reduction

Although Ewald sums are a time honored technique in solid state physics, a considerable amount of effort has recently been devoted to what has come to be called *Lattice reduction* techniques. The basic idea turns around the fact that lattice sums tend to be more practical for $d_\Lambda = 1$ and $d_\Lambda = 3$ than for the grating dimension of $d_\Lambda = 2$.

First one chooses a coordinate system such that a preferred axis (like the z axis) will lie along a given lattice vector. For example $\boldsymbol{a} = (0,0,a)$, and $\boldsymbol{b} = (0,b_2,b_1)$. With this basis the 2D Mono-Layer lattice sum, $S^{ML}_{n,m}$ can then be expressed as a superposition of a Chain sum, $S^{C}_{n,m}$, containing the origin (z-axis), and a superposition of all the chain sums 'above' the central chain (z>0), denoted $S^{ML+}_{n,m}$ or 'below' the central chain, $S^{ML-}_{n,m}$ (z<0). The central chain can be readily be evaluated using one of the techniques described in this chapter, while the integral expression for Hankel functions described in section 6.4.4 allows one to derive efficient expressions for $S^{ML+}_{n,m}$ and $S^{ML-}_{n,m}$.

Lattice reduction can also be applied in the reverse direction, with one expressing the 3D crystalline lattice sum, $S^{L}_{n,m}$, as the superposition of a monolayer sum in the $z = 0$ plane, with sums of all monolayers with $z > 0$, $S^{L+}_{n,m}$ and all monolayers with $z < 0$, $S^{L-}_{n,m}$. There exists efficient techniques for calculating the crystalline lattice sum, $S^{L}_{n,m}$ while one can determine efficient expressions for $S^{L,+}_{n,m}$ and $S^{L,-}_{n,m}$ using again the integral expressions of section 6.4.4. In this *reverse* lattice sum method, the monolayer lattice sum is expressed:

$$S^{ML}_{n,m} = S^{L}_{n,m} - S^{ML,+}_{n,m} - S^{L,-}_{n,m} \ . \tag{6.49}$$

One should that the choice of orientation of the coordinate axis will not be the same in general for different lattice sum techniques, but these differences can be be compensated for by using the rotation matrices of section 6.8.4

Lattice reduction is based on the idea that it can prove numerically efficient to carry out lattice sums for a lattice dimensions other than that desired. To construct a 3D periodic media, we rotate the 2D lattice of the preceding section back to an orientation in the xOy plane with $\boldsymbol{a} = (a,0,0)$, and $\boldsymbol{b} = (b_1,b_2,0)$, and now $\boldsymbol{c} = (0,c_2,c_3)$. The quasi-periodic vector is given by $\boldsymbol{\beta} = (\beta_1,\beta_2,\beta_3)$. The 3D lattice sum can then be written:

$$S^{L}_{n,m} = S^{ML}_{n,m} + S^{L+}_{n,m} + S^{L-}_{n,m} \ . \tag{6.50}$$

The $S^{L+}_{n,m}$ denotes all the $z > 0$ planes, while $S^{L-}_{n,m}$ sums all the $z < 0$ planes.

The lattice reduction technique breaks the sum down into elements which tend to have a decreasing difficulties for divergence. An interest of the lattice reduction technique is that it can be adapted to partial lattices. For instance, large but finite chains, a finite number of infinite chains or finally a finite number of infinite planes.

6.4.2 Plane wave expansion

The expansion of a plane wave in terms of partial waves allows one to transform between partial wave and Fourier transforms. It reads:

$$\begin{aligned} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} &= 4\pi \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\mu=\nu} i^{\nu}\, j_{\nu}(kr)\, Y^{*}_{\nu,\mu}\left(\widehat{\boldsymbol{k}}\right) Y_{\nu,\mu}\left(\widehat{\boldsymbol{r}}\right) \\ &= \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\mu=\nu} p_{\nu,\mu}\, \Psi_{\nu,\mu}(\boldsymbol{r}) \ , \end{aligned} \tag{6.51}$$

where $\Psi_{\nu,\mu}(\boldsymbol{r})$ are the scalar partial wave functions discussed in section 6.2, and $p_{\nu,\mu}$ the coefficients in the development of a scalar plane wave on a partial wave basis i.e. :

$$\Psi_{\nu,\mu}(\boldsymbol{r}) \equiv j_{\nu}(kr)\, Y_{\nu,\mu}\left(\widehat{\boldsymbol{r}}\right) \ , \qquad p_{n,m} = 4\pi i^{n} Y^{*}_{n,m}\left(\widehat{\boldsymbol{k}}\right) \ . \tag{6.52}$$

One can produce an integral expression of $j_n(kr)Y_{n,m}(\widehat{\boldsymbol{r}})$ by multiplying both sides of eq.(6.51) by $Y_{n,m}\left(\widehat{\boldsymbol{k}}\right)$ and integrating over all directions of $\widehat{\boldsymbol{k}}$.

$$\begin{aligned}\int d\Omega_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} Y_{n,m}\left(\widehat{\boldsymbol{k}}\right) &= 4\pi \int d\Omega_{\boldsymbol{k}} \sum_{\nu=0}^{\infty} \sum_{\mu=-\nu}^{\mu=\nu} i^{\nu} j_{\nu}(kr) Y_{\nu,\mu}(\widehat{\boldsymbol{r}}) Y^{*}_{\nu,\mu}\left(\widehat{\boldsymbol{k}}\right) Y_{n,m}\left(\widehat{\boldsymbol{k}}\right) \\ &= 4\pi i^{n} j_n(kr) Y_{n,m}(\widehat{\boldsymbol{r}}) \ . \end{aligned} \tag{6.53}$$

We have thus found that regular partial waves are an angular Fourier transform of the spherical harmonics:

$$\Psi_{\mathscr{J},n,m} \equiv j_n(kr) Y_{n,m}(\widehat{\boldsymbol{r}}) = \frac{1}{4\pi i^n} \int d\Omega_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} Y_{n,m}\left(\widehat{\boldsymbol{k}}\right) \ . \tag{6.54}$$

Likewise, the transverse regular partial waves, $\boldsymbol{\Psi}_{\mathscr{J}}$ can be expressed as an angular Fourier transform of the vector spherical harmonics:

$$\begin{aligned}\boldsymbol{\Psi}_{\mathscr{J},q=1,n,m}(k\boldsymbol{r}) &= \frac{i^{-n}}{4\pi} \int d\Omega_{\boldsymbol{k}}\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \boldsymbol{X}_{n,m}(\widehat{\boldsymbol{k}}) \\ \boldsymbol{\Psi}_{\mathscr{J},q=2,n,m}(k\boldsymbol{r}) &= \frac{i^{1-n}}{4\pi} \int d\Omega_{\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \boldsymbol{Z}_{n,m}(\widehat{\boldsymbol{k}}) \ . \end{aligned} \tag{6.55}$$

6.4.3 Poisson summation formula

The Poisson summation formula is a crucial mathematical tool for evaluating lattice sums. It allows one to pass from a sum over the real lattice vectors to a sum over the reciprocal lattice vectors. Formally, it can be written:

$$\sum_{\boldsymbol{r_j}\in\Lambda}^{\infty} e^{i\boldsymbol{k}\cdot\boldsymbol{r_j}} = \frac{(2\pi)^{d_\Lambda}}{\mathscr{A}_{d_\Lambda}} \sum_{\boldsymbol{p_g}\in\Lambda^*} \delta\left(\boldsymbol{k}-\boldsymbol{p_g}\right) \ , \tag{6.56}$$

where $\mathscr{A}_{d_\Lambda}$ is the "volume" of the reciprocal lattice cell. Since long and short range interactions can both be strong for lattice problems, the Poisson summation formula often does not directly accelerate the lattice sum, but it nevertheless proves invaluable for a number of useful formulas that we will derive in the rest of this chapter.

For the 1-D sum in eq.(6.95), this can be written:

$$\sum_{j=-\infty}^{\infty} e^{i(\boldsymbol{k}+\boldsymbol{\beta})\cdot(\widehat{\boldsymbol{z}}ja)} = \frac{2\pi}{a} \sum_{g=-\infty}^{\infty} \delta\left(k_z+\beta_z-\frac{2\pi}{a}g\right) \ . \tag{6.57}$$

We then write this relation in a dimensionless form:

$$\sum_{j=-\infty}^{\infty} e^{i(\boldsymbol{k}+\boldsymbol{\beta})\cdot(\widehat{\boldsymbol{z}}ja)} = \frac{2\pi}{kai^n} \sum_{g=-\infty}^{\infty} \delta\left(\frac{k_z}{k}+\frac{\beta_z}{k}-g\frac{2\pi}{ka}\right) \ . \tag{6.58}$$

6.4.4 Integral expressions for outgoing partial waves

The Weyl identity expresses the Hankel function of order 0 as an integral of plane waves:

$$h_0(kr) = \frac{1}{2\pi k}\iint_{-\infty}^{\infty} dk_x dk_y \frac{\exp(\pm i\boldsymbol{k}\cdot\boldsymbol{r})}{k_z}$$
$$= \frac{1}{2\pi k}\iint_{-\infty}^{\infty} dk_x dk_y \frac{\exp[\pm i(k_x x + ik_y y + ik_z z)]}{k_z} \qquad z \gtrless 0\,, \tag{6.59}$$

where the plus sign is taken for $z > 0$ and the minus sign is used when $z < 0$. The k_z component is fixed by the constraint that $k_x^2 + k_y^2 + k_z^2 = k^2$, namely $k_z = \sqrt{k^2 - k_x^2 - k_y^2}$. It is interesting to remark that the spherical Bessel function is a superposition of plane waves that are constrained to satisfy $\|\boldsymbol{k}\| = k$. Since the reciprocal space integration in eq.(6.59) is carried out in the xOy plane, it is convenient to define a specific symbol for the wavevector in the xOy plane, $\boldsymbol{K} = k_x\widehat{\boldsymbol{x}} + k_y\widehat{\boldsymbol{y}}$, and the full wavevector is then, $\boldsymbol{k} = \boldsymbol{K} + k_z\widehat{\boldsymbol{z}}$. It is also convenient to define dimensionless factor :

$$\gamma_z \equiv k_z/k = \frac{\sqrt{k^2 - K^2}}{k} \tag{6.60}$$

If we take the position vector $\boldsymbol{r}$ in eq.(6.59) to lie along the z axis, $\boldsymbol{r} = r\widehat{\boldsymbol{z}}$, then we can integrate over the azimuthal angle to obtain a single integral expression for Hankel functions that can be used in lattice sums:

$$h_0(kr) = \frac{1}{k}\int_0^{\infty} dK K \frac{\exp[\pm ik_z r]}{k_z} \qquad z \gtrless 0\,. \tag{6.61}$$

Wittmann pointed out that the above Weyl identity of eq.(6.59) can be generalized to all partial waves of the Hankel function type[33] :

$$\Psi_{\mathscr{H},n,m} \equiv h_n(kr)\,Y_{n,m}(\widehat{\boldsymbol{r}})$$
$$= \frac{i^{-n}}{2\pi k}\iint_{-\infty}^{\infty} dk_x dk_y\,(k_x + ik_y)^m \widetilde{P}_n^m(\gamma_z)\frac{\exp(\pm i(k_x x + k_y y + k_z z))}{k_z} \qquad z \gtrless 0\,. \tag{6.62}$$

If we take $\boldsymbol{r}$ again to lie along the z axis, we find an integral expression for spherical Hankel functions:

$$h_n(kr) = \frac{i^{-n}}{k}\int_0^{\infty} dK K P_n(\gamma_z)\frac{\exp[i\gamma_z kr]}{k_z}\quad . \tag{6.63}$$

The integral relation of eq.(6.62) can also be extended to the outgoing vector partial waves:

$$\begin{aligned}\boldsymbol{\Psi}_{\mathscr{H},q=1,n,m}(k\boldsymbol{r}) &= \tfrac{i^{-n}}{2\pi k}\iint_{-\infty}^{\infty} dk_x dk_y \tfrac{\exp\left(\pm i\left(k_x x + k_y y + k_z z\right)\right)}{k_z}\boldsymbol{X}_{n,m}(\widehat{\boldsymbol{k}}) \\ \boldsymbol{\Psi}_{\mathscr{H},q=2,n,m}(k\boldsymbol{r}) &= \tfrac{i^{1-n}}{2\pi k}\iint_{-\infty}^{\infty} dk_x dk_y \tfrac{\exp\left(\pm i\left(k_x x + k_y y + k_z z\right)\right)}{k_z}\boldsymbol{Z}_{n,m}(\widehat{\boldsymbol{k}})\end{aligned} \qquad z \gtrless 0\quad . \tag{6.64}$$

The Poisson sum rule allows one to express quasi-periodic 2D lattice sum in terms of 2D reciprocal lattice vectors. For the scalar partial waves, one has:

$$\sum_{\boldsymbol{r_j}\in\Lambda} \exp\left(i\boldsymbol{\beta}\cdot\boldsymbol{r_j}\right)\Psi_{\mathscr{H},n,m}\left(k\boldsymbol{r_j}\right) = \sum_{\boldsymbol{p_g}\in\Lambda^*} \frac{2\pi i^{-n}}{k k^+_{\boldsymbol{g},z}\mathscr{A}_2} Y_{n,m}\left(\widehat{\boldsymbol{k}}^{\pm}_{\boldsymbol{g}}\right)\exp\left(i\boldsymbol{k}^{\pm}_{\boldsymbol{g}}\cdot\boldsymbol{r}\right) \qquad z \gtrless 0\,, \quad (6.65)$$

while for the vector partial waves,

$$\sum_{\boldsymbol{r_j}\in\Lambda} \exp\left(i\boldsymbol{\beta}\cdot\boldsymbol{r_j}\right)\boldsymbol{\Psi}_{\mathscr{H},1,n,m}(k\boldsymbol{r_j}) = \sum_{\boldsymbol{p_g}\in\Lambda^*} \frac{2\pi i^{-n}}{k k^+_{\boldsymbol{g},z}\mathscr{A}_2} \boldsymbol{X}_{n,m}\left(\widehat{\boldsymbol{k}}^{\pm}_{\boldsymbol{g}}\right)\exp\left(i\boldsymbol{k}^{\pm}_{\boldsymbol{g}}\cdot\boldsymbol{r}\right)$$

$$\sum_{\boldsymbol{r_j}\in\Lambda} \exp\left(i\boldsymbol{\beta}\cdot\boldsymbol{r_j}\right)\boldsymbol{\Psi}_{\mathscr{H},2,n,m}(k\boldsymbol{r_j}) = \sum_{\boldsymbol{p_g}\in\Lambda^*} \frac{2\pi i^{1-n}}{k k^+_{\boldsymbol{g},z}\mathscr{A}_2} \boldsymbol{Z}_{n,m}\left(\widehat{\boldsymbol{k}}^{\pm}_{\boldsymbol{g}}\right)\exp\left(i\boldsymbol{k}^{\pm}_{\boldsymbol{g}}\cdot\boldsymbol{r}\right) \qquad z \gtrless 0\,. \quad (6.66)$$

In the partial wave lattice sums of eqs.(6.65) and (6.66), the wavevector $\boldsymbol{k}^{\pm}_{\boldsymbol{g}}$ is given by:

$$\boldsymbol{k}^{\pm}_{\boldsymbol{g}} \equiv \left(\boldsymbol{\beta}_{\parallel} + \boldsymbol{p}_{\boldsymbol{g}} \pm \widehat{\boldsymbol{z}}\sqrt{k^2 - \left(\boldsymbol{\beta}_{\parallel} + \boldsymbol{p}_{\boldsymbol{g}}\right)^2}\right)\,, \quad (6.67)$$

and $k^+_{\boldsymbol{g}_z}$ is its z component:

$$k^+_{\boldsymbol{g},z} \equiv \boldsymbol{k}^{\pm}_{\boldsymbol{g}}\cdot\widehat{\boldsymbol{z}} = \sqrt{k^2 - \left(\boldsymbol{\beta}_{\parallel} + \boldsymbol{p}_{\boldsymbol{g}}\right)^2}\,. \quad (6.68)$$

One remarks that for a real Bloch vector $\boldsymbol{\beta}_{\parallel}$, the wavevector $\boldsymbol{k}^{\pm}_{\boldsymbol{g}}$ is real *i.e.* propagative in nature only for those lattice vectors for which

$$k > \left|\left|\boldsymbol{\beta}_{\parallel} + \boldsymbol{p}_{\boldsymbol{g}}\right|\right|\,. \quad (6.69)$$

6.4.5 Partial wave rotation

Let us consider once define a row 'matrix', $\Psi^t_{\mathscr{J}}$, defined as being composed of the regular partial waves defined in eq.(6.5)

$$\Psi^t\left(k\boldsymbol{r}\right) = \left[\mathscr{J}_{0,0}\left(k\boldsymbol{r}\right), \mathscr{J}_{1,-1}\left(k\boldsymbol{r}\right), \mathscr{J}_{1,0}\left(k\boldsymbol{r}\right), ..., \right] \quad (6.70)$$

(or alternatively in terms one of the irregular partial waves in which case the are denoted $\Psi^t_{\mathscr{Y}}$ or $\Psi^t_{\mathscr{H}}$). The arbitrariness of the coordinate system orientation imposes transformation relations amongst the partial waves with the same orbital quantum number n. Let us consider a position M given by the vector $\boldsymbol{r}$ in our chosen coordinate system. We next consider another coordinate system with the same origin, but rotated by the 3 Euler angles, α, β, and γ in which the same point M is now designated by a vector $\boldsymbol{r}'$(*n.b.* $|\boldsymbol{r}'| = |\boldsymbol{r}| = r$). The linear relationship between the row matrix in these 2 coordinate systems is then:

$$\Psi^t\left(k\boldsymbol{r}\right) = \Psi^t\left(k\boldsymbol{r}'\right)\,\mathscr{D}\left(\alpha,\beta,\gamma\right)\,. \quad (6.71)$$

If the rotated coordinate systems is taken such that $\boldsymbol{r}'$ lies along the z axis in the rotated coordinate n this relation takes the form:

$$\Psi^t(k\boldsymbol{r}) = \Psi^t(kr\widehat{\boldsymbol{z}})\,\mathscr{D}(\phi,\theta,0)\ . \tag{6.72}$$

In component form this reads for Hankel function sums:

$$\begin{aligned} h_n(kr)\,Y_{n,m}(\widehat{\boldsymbol{r}}) &= h_n(kr)\,Y_{n0}(0,0)\,\mathscr{D}_{n,0;n,m}(\phi_{\widehat{\boldsymbol{r}}},\theta_{\widehat{\boldsymbol{r}}},0) \\ &= \sqrt{\frac{2n+1}{4\pi}}h_n(kr)\,\mathscr{D}_{n,0;n,m}(\phi_{\widehat{\boldsymbol{r}}},\theta_{\widehat{\boldsymbol{r}}},0)\ , \end{aligned} \tag{6.73}$$

where we used eq.(6.80) for an expression of the $Y_{n,0}(0,0)$.

6.5 Numerical Examples

This section will focuses on infinite chains of particles since this problem provides a relatively simple concrete example of the methods developed in this chapter.

6.5.1 Far and near field response from gratings

As discussed in section 6.3, once the lattice sums have been determined for all the $\Omega_{\boldsymbol{\beta}}$ matrix elements, and the lattice T-matrix of eq.(6.33) obtained, one has ready access to both the far and near field response of the system. However, the quasi-periodic lattice for all the multipole orders must be calculated anew whenever one looks for response to a different quasi-periodicity vector, $\boldsymbol{\beta}$ or wavenumber k (*i.e.* frequency).

6.5.2 Modes for particulate chains

There is considerable interest in calculating and characterizing the 'propagating' modes of periodic chains, gratings, and finite stacks of particulate gratings. For planar surfaces, Greffet has argued[2] that the Leaky-modes can be described by letting either frequency or wavevector be described by a complex number. This idea has recently been employed by several authors for calculating modes in infinite particulate chains where the component of $\boldsymbol{\beta}$ along the chain axis is allowed to be a complex number.[23, 4, 6, 11, 12] The alternative choice of complex frequency seems equally practical for infinite 1-D chains when the particles are lossless. For scatterers composed of dispersive materials however, the complex frequency choice requires an analytical model for permittivity like the Drude or Lorentz models, and all modes become leaky-modes. When analyzing grating systems, it appears simpler to allow for frequency to be the complex parameter, but complex wavevectors remain a viable method. if one defines a complex propagation *vector* in the grating plane.[9].

Typically, one has looked for propagating modes in particulate arrays of sub-wavelength particles metallic particles. There is however an increased interest in high index dielectrics. Due to the complexity of the full multipole approach, most works search for modes in the complex plane have adopted what amounts to be a electric dipole approximation to eq.(6.33).[23] We have recently argued that electric dipole approximation is insufficient in the presence of strong interactions that are provoked by resonances.[23] These results and conclusions are reviewed here.

We adopt the same parameters for a plasmonic chain as Conforti and Guasoni.[6] Namely, we consider an infinite chain of identical 50nm diameter silver particles separated by $d = 75$nm (center-to-center). The system is immersed in a non-magnetic medium with relative permittivity $\varepsilon = 2.25$ ($n = 1.5$).

The figures are plotted with normalized frequencies and wave-vectors:

$$\overline{\omega} \equiv \frac{\omega d}{2\pi c} = \frac{d}{\lambda_v} \qquad \overline{\beta} \equiv \frac{\beta d}{2\pi} \tag{6.74}$$

where λ_v is the vacuum wavelength. The light line for these parameters is given by $\overline{\omega} = \frac{\overline{\beta}}{n_{med}}$. The dispersion relations of the principal propagating modes calculated in the electric dipole approximation are plotted in figure 6.2 (dashed curves). They are then compared with fully converged $n_{max} = 10$ calculations of these dispersion relations (solid line) in this same figure by solving eq.(6.38). The imaginary part of the dispersion relations for dipolar and converged multipole calculations are given in figure 6.2a).

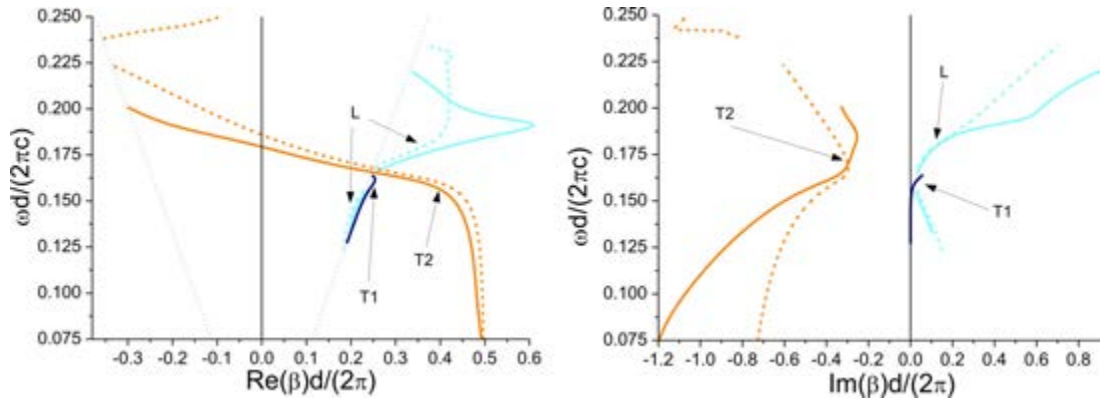


Figure 6.2: Real parts (a) and imaginary parts (b) of the mode dispersion relations in the dipole approximation (dashed curves), and fully converged multipole calculations with $n_{max} = 10$ (full lines). The Longitudinal mode with positive imaginary part is in cyan(gray) the "T1" mode with positive imaginary part is in blue(black line). The "T2" transverse mode with negative imaginary part is in orange(gray). reproduced with permission : http://dx.doi.org/10.1364/JOSAB.29.001012

Figures 6.2a) and 6.2b) merit some commentary. It is immediately clear that the dipole approximation provides a moderately accurate prediction of dispersion relations only over a narrow range of frequencies for which the imaginary part of the propagating wavevector is rather small, and the real part is near the light line. One should also recall that symmetry dictates that if a given value of β corresponds to mode at a given frequency, then by symmetry, $-\beta$ is also a solution to these equations. For the sake of clarity, these symmetric modes are not presented in these figures.

Like Conforti and Guasoni[6], we find a transverse mode, labeled "T2" whose imaginary part of β is opposite in sign with the real part of β. It may prove physically relevant to think of this T2 mode as a backscattering mode, or to interpret this in terms of negative effective index. It is interesting to remark that the T2 mode tends toward the edge of the Brillouin zone at low frequencies.

Our dipole approximation predictions for the longitudinal mode are quite similar to that of ref.[6] wherein the dipole prediction is that the mode "folds back" before reaching the edge of the Brillouin zone. The full multipole calculations on the other hand predict that the longitudinal

mode goes to the edge of the Brillouin zone, and that the "fold back" only occurs after it has gone "beyond" the edge of the Brillouin zone. In our calculations, the "T1" mode is quite close to the light line, and henceforth rather poorly confined by the plasmon chain so its importance in applications seems limited. In our calculations, the dipole approximation for the "T1" mode is quite similar to the multipole solution except that we only found that the full multipole solution predicted both extremities of the T1 mode to lie on the light line.

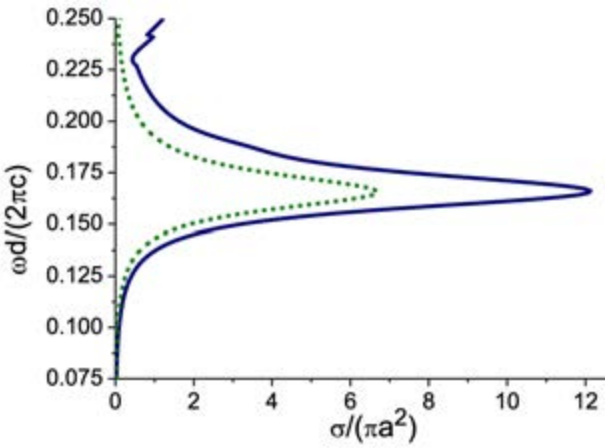


Figure 6.3: Normalized extinction is a solid blue (line) and scattering cross section given by a dashed green line of a silver monomer in terms of frequency ($a = 25$nm). reproduced with permission : http://dx.doi.org/10.1364/JOSAB.29.001012

Due to the system design (sub-wavelength resonant particles) one expected to find significant guiding of modes only in the frequency domains where the scattering cross section of the individual particles is non-negligible. To illustrate this point, we plot the extinction and scattering cross section for an individual particle in the chain in figure 6.3. We remark in particular that near individual particle resonance maxima, all the guided modes of figure 6.2 lie near the light line, and it is also here also that their imaginary parts are smallest. Furthermore, with the exception of the 'backscattering' mode T2, all guided modes apparently cease to exist when one moves sufficiently far away from the scattering resonance frequency.

The reader has probably remarked some strange behavior of the modes in the electric dipole approximation at high frequencies. For instance, at around $\widetilde{\omega} = 0.225$ a "kink" appears in the longitudinal mode, and a spurious T2 solution emerges from the light line. We carried out mode calculations with various multipole cutoffs and found that such kinks and spurious solutions were relatively commonplace (at high or low frequencies) when low numbers of multipoles are used in the simulations and such behavior disappears when higher multipole orders are used. It is also worth remarking that for high order simulations, the $\mathrm{Re}[\beta]$ of the modes terminate at either the light line, or the edge of the Brillouin zone, but modes can terminate at undiscriminating positions in β space when calculations are carried out at low order.

The mode diagrams of figures 6.2a) and 6.2b) were somewhat unconventional since they did not display symmetric, $-\beta$, modes, and allowed the dispersion relation of the longitudinal mode to move outside the Brillouin zone. A more conventional representation of the dispersion relations is given in figure 6.4 which includes the symmetric modes, but only displays modes when $\mathrm{Re}[\beta]$ has positive values lying within Brillouin zone (here we display only the results of multipole calculations). Transverse modes with $\mathrm{Im}[\beta] > 0$ modes are given by solid blue(black) lines, while transverse modes with $\mathrm{Im}[\beta] < 0$ are dashed dashed blue(black) lines. Longitudinal

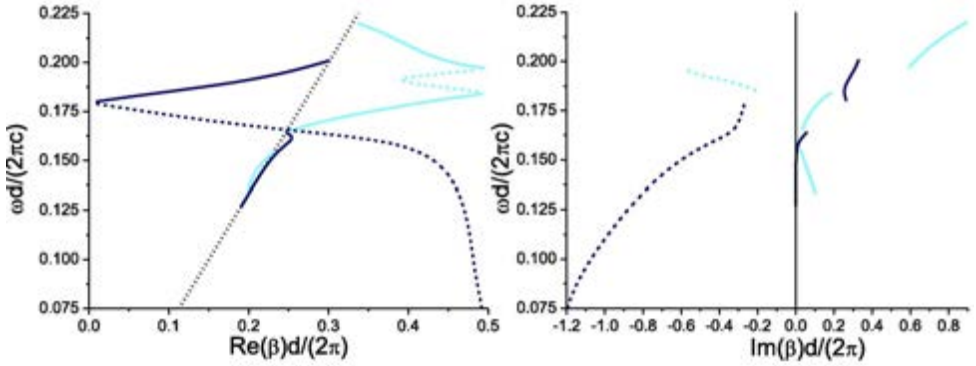


Figure 6.4: Positive and imaginary parts of the dispersion relations in the $Re[\beta] > 0$ part of the Brillouin zone. Transverse modes with: $Im[\beta] > 0$ modes are solid blue(black) lines, while that with $Im[\beta] < 0$ is given by a dashed blue(black) line. Longitudinal modes with $Im[\beta] > 0$ are solid cyan(gray) lines, while those with $Im[\beta] < 0$ is a dashed cyan(gray) line. reproduced with permission : http://dx.doi.org/10.1364/JOSAB.29.001012

modes with $\mathrm{Im}[\beta] > 0$ are given by solid cyan(gray) lines while longitudinal modes with negative $\mathrm{Im}[\beta]$ are in dashed cyan(gray). It is interesting to note that the longitudinal modes extend to the positive edge of the Brillouin zone and that the "fold back" only occurs when $\mathrm{Im}[\beta]$ of the longitudinal mode is negative. One can also remark that transverse modes, T2, with both positive and negative $\mathrm{Im}[\beta]$ exist above the light line, but that their imaginary parts are quite large. Longitudinal modes above the light line also exist at frequencies below the particle resonance maximum, but these modes remain quite close to the light line.

6.6 Chain sums

6.6.1 Hankel function chain sums

A periodic chain of wave scattering is defined by a lattice vector $\boldsymbol{a}$, such that there is an elementary 'scatterer' at all positions $\boldsymbol{r}_j$ i.e.:

$$\boldsymbol{r}_j \equiv j\boldsymbol{a} \qquad j \in \mathbb{Z} \qquad j = -\infty, \ldots -2, -1, 0, 1, 2, \ldots, \infty \ . \tag{6.75}$$

A chain sum for a quasi-periodicity vector $\boldsymbol{\beta}$ is defined:

$$S^{C}_{n,m}(k,a,\beta;\widehat{\boldsymbol{a}}) \equiv S^{C,\mathscr{H}}_{n,m}(k,a,\beta;\widehat{\boldsymbol{a}}) \equiv \sum_{\substack{j\neq 0 \\ j\in\mathbb{Z}}} \mathscr{H}_{n,m}(j\boldsymbol{a})\, e^{ija\boldsymbol{\beta}\cdot\widehat{\boldsymbol{a}}} \ . \tag{6.76}$$

One can remark that the chain sum, $S^{C}_{n,m}$, depends on the amplitude of the lattice vector $a = |\boldsymbol{a}|$, and its direction, and the scalar product between $\boldsymbol{a}$ and another vector $\boldsymbol{\beta}$ which we will call the 'incident' or 'quasi-periodicity' vector. The chain sum in fact only depends on the scalar product between $\boldsymbol{\beta}$ and the periodicity vector:

$$\beta \equiv \boldsymbol{\beta}\cdot\widehat{\boldsymbol{a}} \ . \tag{6.77}$$

One remarks that the direction of $\boldsymbol{a}$ depends on the orientation of the coordinate system. We can take advantage of this fact to define the z axis as the direction of $\boldsymbol{a}$ such that:

$$\boldsymbol{r}_j = ja\widehat{\boldsymbol{z}} \ , \tag{6.78}$$

but one must keep in mind that the expression for $S^C_{n,m}$ is reference frame dependent. In this coordinate system, the chain sum, $S^C_{n,m}(k,a,\beta;\widehat{\mathbf{z}})$, takes the form :

$$\begin{aligned} S^C_{n,m}(k,a,\beta;\widehat{\mathbf{z}}) &\equiv \sum_{\substack{j\neq 0\\ j\in\mathbb{Z}}} \mathscr{H}_{n,m}(ja\widehat{\mathbf{z}})\, e^{i\boldsymbol{\beta}\cdot\widehat{\mathbf{z}}ja} \\ &= \sum_{\substack{j\neq 0\\ j\in\mathbb{Z}}} h_n(k|j|a)\, Y_{n,m}\left(\frac{j}{|j|}\widehat{\mathbf{z}}\right) e^{i\beta aj} \\ &= \delta_{m,0}\lambda_{n,0}\sum_{j=1}^{\infty} h_n(jka)\left[e^{ij\beta a}+(-1)^n e^{-ij\beta a}\right] , \end{aligned} \tag{6.79}$$

where we used the fact that only the $m=0$ scalar spherical harmonics are non-zero at $\theta=0,\pi$:

$$Y_{n,m}(0,0)=\delta_{m,0}\lambda_{n,0}=\sqrt{\frac{2n+1}{4\pi}} \qquad Y_{n,0}(\pi,0)=\delta_{m,0}(-)^n Y_{n,m}(0,0)\ . \tag{6.80}$$

The analytical expressions for the first few Hankel function are:

$$\begin{aligned} h_0(x) &= -\frac{i}{x}e^{ix} \\ h_1(x) &= e^{ix}\left(\frac{-1}{x}-\frac{i}{x^2}\right) \\ h_2(x) &= e^{ix}\left(\frac{i}{x}-\frac{3}{x^2}-\frac{3i}{x^3}\right) , \end{aligned} \tag{6.81}$$

which are readily obtained from the general analytic expression for Hankel functions of arbitrary order:

$$h_n(x)=(-i)^{n+1}\sum_{s=0}^{n}\frac{i^s}{2^s s!}\frac{(n+s)!}{(n-s)!}\frac{e^{ix}}{x^{s+1}}\ . \tag{6.82}$$

6.6.2 Integral technique for Hankel lattice sums

Generalizing the Weyl integral to produce an integral expression for spherical Hankel functions gives us the integral:

$$h_n(kr)=\frac{1}{ki^n}\int_0^{\infty} dK\,K P_n(\gamma_z)\frac{\exp(ik_z r)}{k_z} , \tag{6.83}$$

where we recall that $k_z=\sqrt{k^2-K^2}=k\gamma_z$ and $K=\sqrt{k_x^2+k_y^2}$ is the wavevector component in the xOy plane.

If we try to evaluate this integral numerically, we will encounter problems when we go past the point where $k_z=0$. Since the singularity coming from the k_z denominator lies just above the real axis, we can analytically continue the integration into the fourth quadrant of the complex plane. Any angle will do as long as the resulting line integral is sufficiently far from the positive real axis or the negative imaginary axis. We will generally take an angle of 45° as a reasonable compromise. We will find that the integrand will decrease exponentially for large $|K|$ in the complex plane so that we don't have much problem with the integral extending to infinity.

Thanks to the integral expression for Hankel functions of eq.(6.83), we are now ready to treat an infinite chain sum for a chain oriented along the z axis:

$$\begin{aligned} S^C_{n,m}(k,a,\beta;\widehat{\boldsymbol{z}}) &= \sum_{j\in\mathbb{Z}^*} \exp(i\beta aj)\, h_n(k|j|a)\, Y_{n,m}\left(\frac{j}{|j|}\widehat{\boldsymbol{z}}\right) \\ &= \sum_{j=1}^{\infty} \exp(i\beta aj)\, h_n(kaj)\, Y_{n,m}(0,0) \\ &\qquad + \sum_{j=1}^{\infty} \exp(-i\beta aj)\, h_n(kaj)\, Y_{n,m}(\pi,0) \\ &= \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}}\left[\sum_{j=1}^{\infty} \exp(i\beta aj)\, h_n(kaj) + (-)^n \sum_{j=1}^{\infty} \exp(-i\beta aj)\, h_n(kaj)\right] \end{aligned} \tag{6.84}$$

where we used:

$$Y_{n,m}(0,0) = \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}} P_n(1)\ , \qquad Y_{n,m}(\pi,0) = \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}} P_n(-1)\ , \tag{6.85}$$

and

$$P_n(1) = 1\ , \qquad P_n(-1) = (-1)^n\ . \tag{6.86}$$

Using the integral relation of eq.(6.83), we have:

$$\begin{aligned} S^C_{n,m}(k,a,\beta;\widehat{\boldsymbol{z}}) &= \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}}\frac{1}{i^n k}\int_0^{\infty} dK\, K\, \frac{P_n(k_z/k)}{k_z} \\ &\quad \times \left[\sum_{j=1}^{\infty} \exp\left[i\left(k_z+\beta\right)jd\right] + (-)^n \sum_{j=1}^{\infty} \exp\left[i\left(k_z-\beta\right)ja\right]\right]\ . \end{aligned} \tag{6.87}$$

We have finally an integral expression for the chain sums:

$$\begin{aligned} S^C_{n,m}(k,a,\beta;\widehat{\boldsymbol{z}}) &= \sum_{j\in\mathbb{Z}^*} \exp(i\beta jd)\, h_n\left(kr_j\right) Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) \\ &= \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}}\frac{1}{ki^n}\int_0^{\infty} dK\, K\frac{P_n(k_z)}{k_z} \\ &\quad \times \left[\frac{1}{\exp\left[-i\left(k_z+\beta\right)a\right]-1} + \frac{(-)^n}{\exp\left[-i\left(k_z-\beta\right)a\right]-1}\right]\ . \end{aligned} \tag{6.88}$$

6.6.3 Polylog approach to Hankel chain sums

Inspection of eqs.(6.84) and (6.82) shows that all terms in the chain sum can be expressed in terms of polylogarithm functions which are defined by[1]:

$$\mathrm{Li}_n(z) = \sum_{j=1}^{\infty} \frac{z^j}{j^n}\ . \tag{6.89}$$

The chain sum expressed in terms of polylogarithms is then:

$$S^{C}_{n,m}(k,a,\beta;\widehat{\boldsymbol{z}}) = \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}}\sum_{s=0}^{n}\left[\left((-i)^{n+1}\frac{i^s}{2^s s!}\frac{(n+s)!}{(n-s)!}\right)\right.$$
$$\left.\times\frac{(\mathrm{Li}_{s+1}\exp[i(k+\beta)a]+(-)^n\mathrm{Li}_{s+1}\exp[i(k-\beta)a])}{(ka)^{s+1}}\right]. \tag{6.90}$$

This was the chain sum for outgoing Hankel functions, but we will also sometimes be interested in incoming Hankel functions, or Bessel functions of the fourth kind. These are expressed:

$$h^{(4)}_n(x) \equiv h^{-}_n(x) = j_n(x) - iy_n(x)\ , \tag{6.91}$$

and their chain sums are:

$$S^{C}_{n,m}(k,a,\beta;\widehat{\boldsymbol{z}}) = \delta_{m,0}\sqrt{\frac{2n+1}{4\pi}}\sum_{s=0}^{n}\left[\left((-i)^{n+1}\frac{i^s}{2^s s!}\frac{(n+s)!}{(n-s)!}\right)\right.$$
$$\left.\times\frac{(\mathrm{Li}_{s+1}\exp[-i(k-\beta)a]+(-)^n\mathrm{Li}_{s+1}\exp[-i(k+\beta)a])}{(ka)^{s+1}}\right]. \tag{6.92}$$

6.6.4 Bessel function chain sums

Although fully analytic expressions for Hankel function chain and lattice sums do not seem to exist currently, the Bessel functions lattice and chain sums do have analytic expressions. These Bessel function sums are useful in their own right for certain applications:

$$\begin{aligned} S^{C,\mathscr{J}}_{n}(k,a,\beta;\widehat{\boldsymbol{z}}) &= \sum_{j=-\infty,j\neq 0}^{\infty} Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) j_n(jka)\, e^{ij\beta a} \\ &= \sum_{j=-\infty}^{\infty} Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) j_n(jka)\, e^{ij\beta a} - \sum_{j=-\infty}^{\infty} Y_{0,0}\left(\widehat{\boldsymbol{r}}_0\right) j_0(jka) \\ &= \sum_{j=-\infty}^{\infty} Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) j_n(jka)\, e^{ij\beta a} - \frac{1}{\sqrt{4\pi}}\delta_{n,0}\ . \end{aligned} \tag{6.93}$$

Using the integral expression for $Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) j_n(jka)$ as an integral over the directions a wavenumber $\widehat{\boldsymbol{k}}$ as derived in eq.(6.54) allows us to write:

$$j_n\left(kR_j\right) Y_{n,m}\left(\widehat{\boldsymbol{r}}_j\right) e^{i\boldsymbol{\beta}\cdot(\widehat{\boldsymbol{z}}ja)} = \frac{1}{4\pi i^n}\int Y_{n,m}\left(\widehat{\boldsymbol{k}}\right) e^{i\boldsymbol{\beta}\cdot\widehat{\boldsymbol{z}}ja} e^{i\boldsymbol{k}\cdot\widehat{\boldsymbol{z}}ja} d\Omega_{\boldsymbol{k}}\ . \tag{6.94}$$

The lattice sum of the Bessel type then can be written:

$$S^{C,\mathscr{J}}_{n}(k,a,\beta;\widehat{\boldsymbol{z}}) = \frac{1}{4\pi i^n}\int Y_{n,m}\left(\widehat{\boldsymbol{k}}\right)\left[\sum_{j=-\infty}^{\infty} e^{i(\boldsymbol{k}+\boldsymbol{\beta})\cdot(\widehat{\boldsymbol{z}}ja)}\right] d\Omega_{\boldsymbol{k}} - \frac{1}{\sqrt{4\pi}}\delta_{n,0}\ . \tag{6.95}$$

At this point, one invokes the Poisson summation formula which can be written formally as:

$$\sum_{\boldsymbol{r_j}\in\Lambda}^{\infty} e^{i\boldsymbol{k}\cdot\boldsymbol{r}_j} = \frac{(2\pi)^{d_\Lambda}}{\mathscr{A}} \sum_{\boldsymbol{p}_g\in\Lambda^*} \delta\left(\boldsymbol{k}-\boldsymbol{p}_g\right) \ , \tag{6.96}$$

where $\mathscr{A}$ is the "volume" of the reciprocal cell. For the 1-D sum in eq.(6.95), this can be written:

$$\sum_{j=-\infty}^{\infty} e^{i(\boldsymbol{k}+\boldsymbol{\beta})\cdot(\widehat{\boldsymbol{z}}ja)} = \frac{2\pi}{a} \sum_{g=-\infty}^{\infty} \delta\left(k_z+\beta_z-\frac{2\pi}{a}g\right) \ . \tag{6.97}$$

We then write this relation in a dimensionless form:

$$\sum_{j=-\infty}^{\infty} e^{i(\boldsymbol{k}+\boldsymbol{\beta})\cdot(\widehat{\boldsymbol{z}}ja)} = \frac{2\pi}{kai^n} \sum_{g=-\infty}^{\infty} \delta\left(\frac{k_z}{k}+\frac{\beta_z}{k}-g\frac{2\pi}{ka}\right) \ . \tag{6.98}$$

Putting this relation into the $\widehat{\boldsymbol{k}}$ integral of eq.(6.95), we then obtain a finite sum expression for $S_n^{C,\mathscr{J}}$:

$$S_n^{C,\mathscr{J}}\left(k,a,\beta;\widehat{\boldsymbol{z}}\right) = -\frac{1}{\sqrt{4\pi}}\delta_{n,0} + \frac{\pi i^n}{ka} \sum_{g=g_{\min}}^{g_{\max}} Y_{n0}\left(\cos\beta_{z,q}\right) \ , \tag{6.99}$$

where since -1 $< k_z/k <$ 1 we only sum over those values of g for which

$$-1 < \Re\left[\frac{\beta_z a+2\pi g}{ka}\right] < 1 \ . \tag{6.100}$$

The values $g_{\min}$ and $g_{\max}$ are:

$$g_{\min} = \left(-\frac{\beta_z a-ka}{2\pi}\right)+1 \qquad g_{\max} = -\frac{\beta_z a+ka}{2\pi} \ . \tag{6.101}$$

The angle $\cos\beta_{z,g}$ in eq.(6.99) is given by:

$$\cos\beta_{z,g} \equiv \frac{\beta_z a+2\pi g}{ka} = \Re\left[\beta_z/k\right]+i\Im\left[\beta_z/k\right]+g\frac{2\pi}{ka} \ , \tag{6.102}$$

where we used the parity relation:

$$Y_{n,m}\left(-\widehat{\boldsymbol{r}}\right) = (-1)^n Y_{n,m}\left(-\widehat{\boldsymbol{r}}\right) \ . \tag{6.103}$$

We recall that the sin and cosines for a complex angle, $\theta_k = \theta' + \theta''$, are given by:

$$\begin{aligned}\cos\theta_k &= \frac{e^{i\theta'}e^{-\theta''}+e^{-i\theta'}e^{\theta''}}{2} \\ &= \cos\theta'\cosh\theta'' - i\sin\theta'\sinh\theta'' \ ,\end{aligned} \tag{6.104}$$

and

$$\begin{aligned}\sin\theta_k &= \frac{e^{i\theta'}e^{-\theta''}-e^{-i\theta'}e^{\theta''}}{2i} \\ &= \sin\theta'\cosh\theta''+i\cos\theta'\sinh\theta'' \ .\end{aligned} \tag{6.105}$$

6.6.5 Chain sum rotation

The chain sums expressions given in eqs.(6.88), (6.90), and (6.99) all took advantage of the facilities presented by orienting the chain of particles along the z axis. When performing lattice reduction techniques, it is necessary to have chain sums in other orientations. The chain sum is obtained by applying:

$$Y_{n,m}(\widehat{\boldsymbol{r}}) = Y_{n,0}(\widehat{\boldsymbol{z}})\,\mathscr{D}^{(n)}_{0,m}(\theta_{\widehat{\boldsymbol{r}}},\phi_{\widehat{\boldsymbol{r}}}) = \sqrt{\frac{2n+1}{4\pi}}\mathscr{D}^{(n)}_{0,m}(\theta_{\widehat{\boldsymbol{r}}},\phi_{\widehat{\boldsymbol{r}}})\ , \tag{6.106}$$

which just translates the relation derived in Edmonds (eq.(4.1.25) page 59) that:

$$\mathscr{D}^{(n)}_{0,m}(\theta,\phi) = \sqrt{\frac{4\pi}{2n+1}}Y_{n,m}(\theta,\phi)\ . \tag{6.107}$$

Thus is trivial to write chain sums of any type ($\mathscr{J}$,$\mathscr{H}$,$\mathscr{Y}$) in an arbitrary orientation, $\widehat{\boldsymbol{r}}$, in terms of the chain sum in the direction $\widehat{\boldsymbol{z}}$ by the simple relation:

$$S^{C}_{n,m}(\beta;\widehat{\boldsymbol{r}}) = S^{C}_{n}(\beta;\widehat{\boldsymbol{z}})\sqrt{\frac{4\pi}{2n+1}}Y_{n,m}(\theta_{\widehat{\boldsymbol{r}}},\phi_{\widehat{\boldsymbol{r}}})\ . \tag{6.108}$$

An orientation along the x axis is for example:

$$S^{C}_{n,m}(\beta;\widehat{\boldsymbol{x}}) = S^{C}_{n}(\beta;\widehat{\boldsymbol{z}})\sqrt{\frac{4\pi}{2n+1}}Y_{n,m}\left(\frac{\pi}{2},0\right)\ , \tag{6.109}$$

is useful when carrying out a monolayer sum in the next section. An orientation along the y axis is for example:

$$S^{C}_{n,m}(\beta;\widehat{\boldsymbol{y}}) = S^{C}_{n}(\beta;\widehat{\boldsymbol{z}})\sqrt{\frac{4\pi}{2n+1}}Y_{n,m}\left(\frac{\pi}{2},\frac{\pi}{2}\right)\ . \tag{6.110}$$

6.7 Grating lattice sums

A 2D periodic media is characterized by two basic lattice vectors $\boldsymbol{a}$, $\boldsymbol{b}$. Although, we want to describe a system with lattice vectors $\boldsymbol{a} = (a,0,0)$, and $\boldsymbol{b} = (b_1,b_2,0)$, we are going to work in a rotated coordinate systems in which the lattice will be placed in the xOy plane.

6.7.1 Integral technique

For the integral technique, it is useful to adopt a coordinate system where the $\boldsymbol{a}$ lattice vector lies along the y axis. In this coordinate system, the basis vectors are $\boldsymbol{a} = (0,a,0)$, $\boldsymbol{b} = (0,b_2,b_1)$, then lattice sites are given by:

$$\boldsymbol{r}_{\boldsymbol{j}=(j_a,j_b)} = j_a\boldsymbol{a} + j_b\boldsymbol{b} = (0, j_a a + j_b b_2, j_b b_1)\ . \tag{6.111}$$

In this case, fixing $j_b = 0$ and summing over j_a corresponds to a chain sum along the y axis, and $j_b \lessgtr 0$, corresponds to a term in the $z \lessgtr 0$ half plane respectively.

The Mono-Layer (ML) lattice sum, $S^{ML}_{n,m}$, can be written:

$$S^{ML}_{n,m} = S^{C}_{n,m}(\beta_1;\widehat{\boldsymbol{y}}) + S^{ML+}_{n,m} + S^{ML-}_{n,m}\ , \tag{6.112}$$

where $S^{C}_{n,m}(\beta_1;\widehat{\boldsymbol{y}})$ is the chain sum along the y axis, and $S^{ML}_{n,m}$ is the sum of all the sites with $z \neq 0$ The lattice sum for all $j_b > 0$ ($z > 0$) sites can be expressed as an integral:

$$S^{ML+}_{n,m} = -\frac{(-)^m}{i^n ka} \sum_{g=-\infty}^{\infty} \int_0^{\infty} \frac{dk_x}{k_z} \widetilde{P}^{(m)}_n(\gamma) \left[(k_g - ik_x)^m + (k_g + ik_x)^m\right] \times \frac{1}{1 - \exp\left\{-i\left[b_2(\beta_2 - k_g) + b_1(\beta_1 + \gamma)\right]\right\}} , \tag{6.113}$$

while the lattice sum for all $j_b < 0$, can be expressed:

$$S^{ML-}_{n,m} = -\frac{1}{(-i)^n ka} \sum_{g=-\infty}^{\infty} \int_0^{\infty} \frac{dk_x}{k_z} \widetilde{P}^{(m)}_n(\gamma) \left[(k_g - ik_x)^m + (k_g + ik_x)^m\right] \times \frac{1}{1 - \exp\left\{-i\left[b_2(-\beta_2 + k_g) + b_1(-\beta_1 + \gamma)\right]\right\}} . \tag{6.114}$$

In both of these expressions, we defined k_g as the reciprocal lattice vector along the y axis:

$$k_g \equiv \beta_2 + 2\pi g/a , \tag{6.115}$$

and γ is the reciprocal lattice vector along the z axis:

$$\gamma \equiv \sqrt{1 - k_g^2 - k_x^2} . \tag{6.116}$$

At the end of this calculation, one should keep in mind that lattice sum was carried out in a system where the lattice was in the yOz plane. One can obtain the expression for the lattice sum in the xOy plane by rotating the lattice sums by 90° around the y axis in a clockwise manner, and then 90° around the new y' axis, and finally 90° around the new z'' axis.

6.7.2 Modified Bessel function sums

Although the integral technique is rather efficient, one may prefer to obtain do a little more analytic work and obtain the lattice sum in a manner which takes the from of a lattice sum in a 2D host space. As for the integral it proves convenient to place the lattice in the yOz plane, but this time, one places the $\boldsymbol{a}$ lattice vector along the z axis, so that the lattice vectors can be written:

$$\boldsymbol{a} = (0,0,a), \qquad \text{and} \qquad \boldsymbol{b} = (0,b_2,b_1) .$$

The lattice sites in this system can be expressed:

$$\boldsymbol{r}_{\boldsymbol{j}=(j_a,j_b)} = j_a\boldsymbol{a} + j_b\boldsymbol{b} = (0, j_b b_2, j_a a + j_b b_1) = j_b b_2 \widehat{\boldsymbol{y}} + (j_a a + j_b b_1)\widehat{\boldsymbol{z}} . \tag{6.117}$$

The reciprocal lattice is given by:

$$\boldsymbol{K}_{\boldsymbol{g}=(g_a,g_b)} = g_a\widetilde{\boldsymbol{a}} + g_b\widetilde{\boldsymbol{b}} , \tag{6.118}$$

where the reciprocal lattice vectors are:

$$\widetilde{\boldsymbol{a}} = \frac{1}{ab_2}(0,-b_1,b_2) \qquad \widetilde{\boldsymbol{b}} = \left(0, \frac{1}{b_2}, 0\right) . \tag{6.119}$$

This time, lattice reduction is performed by carrying out the lattice sum on the z axis of the working coordinate system which is to say that one sets $j_b = 0$, and sums over all j_a. The lattice sum is then achieved by:

$$S_{n,m}^{ML} = S_{n,m}^{C}(\beta_1;\widehat{\boldsymbol{z}}) + S_{n,m}^{ML,+} + S_{n,m}^{ML,-} \ , \tag{6.120}$$

where $S_{n,m}^{ML,\pm}$ is the sum of all sites except those along the z axis:

$$S_{n,m}^{ML,\pm} = \sum_{j_b\in\mathbb{Z}^*} e^{ij_b(\beta_1 b_1+\beta_2 b_2)} \sum_{j_a=-\infty}^{\infty} e^{ij_a\beta_1 a}\mathscr{H}_{n,m}\left(k\boldsymbol{r_j}\right) \ . \tag{6.121}$$

Since $j_b \neq 0$ in this sum and all lattice sites are in the yOz plane, the azimuthal angle of $\boldsymbol{r_j}$ is either $\pi/2$ or $-\pi/2$ for $j_b > 0$ or $j_b < 0$ respectively. This allows us to conclude in this plane, $\mathscr{H}_{n,m}$ doesn't depend on the sign of m:

$$\mathscr{H}_{n,-m}\left(k\boldsymbol{r_j}\right) = \mathscr{H}_{n,m}\left(k\boldsymbol{r_j}\right). \tag{6.122}$$

We then appeal to an integral representation:

$$h_n(kr)\overline{P}_n^m(\cos\theta) = \frac{(-i)^{n+1}}{\pi}\int_{-\infty}^{\infty} e^{ikzt}K_m(-ik\rho\gamma(t))\overline{P}_n^m(t)\,dt \tag{6.123}$$

$$= \frac{(-i)^{n-m}}{\pi}\int_{-\infty}^{\infty} e^{ikzt}H_m(k\rho\gamma(t))\overline{P}_n^m(t)\,dt \ , \tag{6.124}$$

where $z = r\cos\theta$, $k\rho = \sqrt{(kr)^2-(kz)^2} > 0$, and $K_m(z)$ is a modified Bessel function defined by:

$$K_m(z) \equiv i^{m+1}H_m(iz) \ , \tag{6.125}$$

and finally $\gamma(t)$ is defined such that:

$$\gamma(t) = \begin{cases} i\sqrt{t^2-1} & |t|\geq 1 \\ \sqrt{1-t^2} & t<1 \end{cases} \ . \tag{6.126}$$

The modified Bessel functions, $K_m(z)$. The j_a sum can then be written:

$$\sum_{j_a=-\infty}^{\infty} e^{ij_a\beta_1 a}\mathscr{H}_{n,m}\left(k\boldsymbol{r_j}\right)$$

$$= \frac{(-i)^n}{\pi}(-)^m[\mathrm{sgn}(j_b)]^m \sum_{j_a=-\infty}^{\infty} e^{ij_a\beta_1 a}\int_{-\infty}^{\infty} e^{ik(j_a a+j_b b_1)t}H_m(k\rho\gamma(t))\overline{P}_n^m(t)\,dt \tag{6.127}$$

$$= \frac{(-i)^n}{\pi}(-)^m[\mathrm{sgn}(j_b)]^m \int_{-\infty}^{\infty}\sum_{j_a=-\infty}^{\infty} e^{ij_a(\beta_1+kt)a}e^{ikj_b b_1 t}H_m(k\rho\gamma(t))\overline{P}_n^m(t)\,dt \ . \tag{6.128}$$

Using the 1D Poisson sum formula, we have:

$$\sum_{j_a=-\infty}^{\infty} e^{ij_a(kt+\beta_1)a} = \frac{2\pi}{a}\sum_{g=-\infty}^{\infty}\delta\left(kt+\beta_1+g\frac{2\pi}{a}\right)$$

$$= \frac{2\pi}{ka}\sum_{g=-\infty}^{\infty}\delta\left(t+\frac{\beta_1}{k}+g\frac{2\pi}{ka}\right) \ . \tag{6.129}$$

$$\begin{aligned}
&\sum_{j_a=-\infty}^{\infty} e^{ij_a\beta_1 a}\mathscr{H}_{n,m}\left(k\boldsymbol{r_j}\right) \\
&\qquad = \frac{(-i)^n}{\pi}(-)^m\left[\mathrm{sgn}\left(j_b\right)\right]^m\int_{-\infty}^{\infty} e^{ikj_b b_1 t}\sum_{j_a=-\infty}^{\infty} e^{ij_a(\beta_1+kt)a}H_m\left(k\rho\gamma_z(t)\right)\overline{P}_n^m(t)\,dt \\
&\qquad = \frac{2(-i)^n}{ka}(-)^m\left[\mathrm{sgn}\left(j_b\right)\right]^m\sum_{g=-\infty}^{\infty} e^{-i\beta_{1p}j_b b_1}H_m\left(kb_2\left|j_b\right|\gamma_g\right)\overline{P}_n^m\left(-\overline{\beta}_{1,g}\right) \\
&\qquad = \frac{2i^n}{ka}\left[\mathrm{sgn}\left(j_b\right)\right]^m\sum_{g=-\infty}^{\infty} e^{-i\beta_{1,g}j_b b_1}H_m\left(kb_2\left|j_b\right|\gamma_g\right)\overline{P}_n^m\left(\overline{\beta}_{1,g}\right)\ , \qquad (6.130)
\end{aligned}$$

where $\beta_{1,g}$ and $\overline{\beta}_{1,g}$ are defined:

$$\beta_{1,g}\equiv\beta_1+g\frac{2\pi}{a}\qquad \overline{\beta}_{1,g}\equiv\frac{\beta_1}{k}+g\frac{2\pi}{ka} \qquad (6.131)$$

and

$$\gamma_g\equiv\gamma\left(\overline{\beta}_{1,g}\right)\ . \qquad (6.132)$$

We have therefore the monolayer sum:

$$\begin{aligned}
S_{n,m}^{ML,\pm} &= \frac{2i^n}{ka}\sum_{g=-\infty}^{\infty}\overline{P}_n^m\left(\overline{\beta}_{1,g}\right)\sum_{j_b\in\mathbb{Z}^*} e^{ij_b(\beta_1-\beta_{1,g})b_1}e^{ij_b\beta_2 b_2}\left[\mathrm{sgn}\left(j_b\right)\right]^m H_m\left(kb_2\left|j_b\right|\gamma_g\right) \\
&= \frac{2i^n}{ka}\sum_{g=-\infty}^{\infty}\overline{P}_n^m\left(\overline{\beta}_{1,g}\right)\sum_{j_b=1}^{\infty}\left[e^{ij_b w_g}+(-)^m e^{-ij_b w_g}\right]H_m\left(kb_2 j_b\gamma_g\right) \\
&= \frac{2i^n}{ka}\sum_{g=-\infty}^{\infty}\overline{P}_n^m\left(\overline{\beta}_{1,g}\right)S_m\left(w_g,kb_2\gamma_g\right)\ , \qquad (6.133)
\end{aligned}$$

where we defined:

$$w_g\equiv\beta_2 b_2-\frac{2p\pi b_1}{a}\ . \qquad (6.134)$$

One can rotate the lattice sum to put it in the desired coordinate system in the *xOy* plane. This can be obtained simply by a rotation of 90° about the *y* axis. The expression of eq.(6.133) is still in the form of the 2 double infinite sums like our initial expression. However, only a finite number of g values will correspond to propagating modes, i.e. $\left|\overline{\beta}_{1,g}\right|<1$, and the other values for g correspond to evanescent modes and are exponentially convergent. therefore infinite series sums. The j_b sum in eq.(6.133) is known as a Schlömilch series, and it can be expressed as a finite sum of Bernoulli polynomials.

6.7.3 Schlömilch series

The Schlömilch series can be expressed

$$S_m(\lambda,\mu)\equiv\sum_{j=1}^{\infty}\left[e^{i\lambda j}+(-)^m e^{-i\lambda j}\right]H_m(\mu j)\ . \qquad (6.135)$$

The zero order sum is

$$S_0(\lambda,\mu) = -1 - \frac{2i}{\pi}\left(C + \log\frac{\mu}{4\pi}\right) + \frac{2}{\Theta_0} + \sum_{g\in\mathbb{Z}^*}\left(\frac{2}{\Theta_g} + \frac{i}{\pi|g|}\right) , \tag{6.136}$$

where $C \simeq 0.5772$ is Euler's constant and

$$\Theta_g = \left(\mu^2 - \lambda_g^2\right)^{1/2} \qquad \lambda_g = \lambda + 2g\pi . \tag{6.137}$$

6.8 Addition theorem and Rotation matrices

6.8.1 Scalar spherical harmonics

The scalar spherical harmonics, $Y_{n,m}(\theta,\phi)$, are expressed in terms of the associated Legendre functions $P_n^m(\cos\theta)$ [7] :

$$\begin{aligned} Y_{n,m}(\theta,\phi) &= \left[\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}\right]^{\frac{1}{2}} P_n^m(\cos\theta)\exp(im\phi) \\ &\equiv \overline{P}_n^m(\cos\theta)\exp(im\phi) , \end{aligned} \tag{6.138}$$

where in the second line we have introduced the normalized associated Legendre functions, $\overline{P}_n^m(\cos\theta_k) \equiv \lambda_{n,m}P_n^m(\cos\theta_k)$, where the $\lambda_{n,m}$ normalization factor is defined:

$$\lambda_{n,m} \equiv \left[\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}\right]^{\frac{1}{2}} . \tag{6.139}$$

These scalar spherical harmonics are normalized with respect to an integration over the solid angles :

$$\begin{aligned} \int_0^{4\pi} d\Omega\, Y^*_{\nu,\mu}(\theta,\phi)\, Y_{n,m}(\theta,\phi) &\equiv (-1)^{\mu}\int_0^{\pi}\sin\theta d\theta\int_0^{2\pi} d\phi\, Y_{\nu,-\mu}(\theta,\phi)\, Y_{n,m}(\theta,\phi) \\ &= \delta_{n,\nu}\delta_{m,\mu} . \end{aligned} \tag{6.140}$$

In principal, the Legendre polynomials, $P_n(x) = P_n^0(x)$, can be obtained from Rodrigues' formula:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left(x^2-1\right)^n , \tag{6.141}$$

but in practice we will calculate them with recurrence relations. Likewise, the associated Legendre functions could be obtained for from the expression:

$$P_n^m(x) = (-1)^m\left(1-x^2\right)^{m/2}\frac{d^m}{dx^m}P_n(x) . \tag{6.142}$$

Their calculation is simplified by noting that the *normalized* associated Legendre functions have the convenient parity property that:

$$\overline{P}_n^{-m}(x) = (-1)^m\overline{P}_n^m(x) . \tag{6.143}$$

There are alternative ways of calculating the scalar spherical harmonics that are better for formulating lattice sums and reflections from a physical interface. In lattice sums and reflections from surfaces, the spherical harmonics will be evaluated in terms of the direction of the incident or reflected wavevectors, $\widehat{\boldsymbol{k}}$:

$$Y_{n,m}(\theta_k,\phi_k) = Y_{n,m}\left(\widehat{\boldsymbol{k}}\right) = Y_{n,m}(\boldsymbol{k}/k) = Y_{n,m}(k_x/k, k_y/k, k_z/k) \ , \tag{6.144}$$

where we recall that:

$$\begin{aligned} \frac{k_z}{k} &= \cos\theta_k \\ k_x/k &= \sin\theta_k \cos\phi_k \\ k_y/k &= \sin\theta_k \sin\phi_k \ , \end{aligned} \tag{6.145}$$

and we keep in mind that $\overline{P}_n^m$ are functions of $\cos\theta_k = k_z/k$.

Since $x = \cos\theta$, and the $P_n(x)$ are polynomials in x, the $\frac{d^m}{dx^m}P_n(x)$ are functions of $\cos\theta$. The factor $\left(1-x^2\right)^{m/2}$ corresponds to $\sin^m\theta_k$ with no ambiguity in sign since $\mathrm{Re}\{\theta_k\}$ is $\in (0,\pi)$. One should remark that the $\left(1-x^2\right)^{m/2}$ is non-polynomial so that is why one refers to them as associated Legendre *functions*. For applications involving reciprocal space and/or integrations in the complex plane it proves useful to explicitly extract this factor, and define associated Legendre *polynomials*, which we shall denote, $\widetilde{P}_n^m$ (not to be confused with the normalized associated Legendre functions).

For positive m we have then:

$$\begin{aligned} Y_{n,m}(k_x/k, k_y/k, k_z/k) &= \overline{P}_n^m(\cos\theta_k)\exp(im\phi_k) \\ &= \lambda_{n,m}(-1)^m \sin^m\theta_k (\cos\phi_k + i\sin\phi_k)^m \frac{d^m}{dx^m}P_n(\frac{k_z}{k}) \\ &= (\sin\theta_k\cos\phi_k + i\sin\theta_k\sin\phi_k)^m (-1)^m \lambda_{n,m}\frac{d^m}{dx^m}P_n(\frac{k_z}{k}) \\ &= (k_x/k + ik_y/k)^m (-1)^m \lambda_{n,m}\frac{d^m}{dx^m}P_n(\frac{k_z}{k}) \\ &= (k_x/k + ik_y/k)^m \widetilde{P}_n^m(\frac{k_z}{k}) = \left(\frac{K}{k}\right)^{|m|} \exp(im\phi_k)\widetilde{P}_n^m(\frac{k_z}{k}) \ . \end{aligned} \tag{6.146}$$

where we have defined the normalized associated Legendre *polynomials*, $\widetilde{P}_n^m$, such that:

$$\widetilde{P}_n^m(\frac{k_z}{k}) \equiv (-1)^m \lambda_{n,m}\frac{d^m}{dx^m}P_n(\frac{k_z}{k}) \ . \tag{6.147}$$

The parameter,

$$\boldsymbol{K} \equiv k_x\widehat{\boldsymbol{x}} + k_y\widehat{\boldsymbol{y}} \ , \tag{6.148}$$

corresponds to the momentum space vector in the x-y plane.

The wave vector components k_x, k_y, and k_z are related to the *possibly complex* angles, θ_k and ϕ_k, via the relations:

$$\begin{aligned} k_x &= k\sin\theta_k\cos\phi_k \\ k_y &= k\sin\theta_k\sin\phi_k \\ k_z^2 &= k^2 - K^2 = k^2\cos^2\theta_k \ , \end{aligned} \tag{6.149}$$

This relation of eq.(6.146) for $Y_{n,m}$ can be extended to negative m by writing :

$$Y_{n,m}(k_x/k, k_y/k, k_z/k) = \left(\frac{K}{k}\right)^{|m|} \exp(im\phi_k)\, \widetilde{P}_n^m(\frac{k_z}{k}) \qquad m \gtrless 0\ , \tag{6.150}$$

as long as we define $\widetilde{P}_n^{-m}$ such that :

$$\widetilde{P}_n^{-m} \equiv (-1)^m \widetilde{P}_n^m\ . \tag{6.151}$$

The objective of the above procedure was to define $\widetilde{P}_n^m(x)$ that are always polynomials of x for both positive and negative m. This is in contrast to the associated Legendre functions $\overline{P}_n^m(x)$ which are not polynomials in terms of x.

6.8.2 Translation-addition theorem for scalar partial waves

Let us consider a point M in a system using spherical coordinates. We consider a second system of spherical coordinates centered on the position $\boldsymbol{r}_0$. The position of M in this second system centered on $\boldsymbol{r}_0$ is:

$$\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{r}_0\ . \tag{6.152}$$

We take the usual convention of outgoing scalar partial waves as products of spherical Hankel function and scalar spherical harmonics:

$$\Psi_{\mathscr{H},n,m}(k\boldsymbol{r}) \equiv \Psi^{(3)}_{n,m}(k\boldsymbol{r}) \equiv h_n(kr)\, Y_{n,m}(\theta,\phi)\ , \tag{6.153}$$

while the regular scalar partial waves replace the spherical Hankel functions with spherical Bessel functions:

$$\Psi_{\mathscr{J},n,m}(k\boldsymbol{r}) \equiv \Psi^{(1)}_{n,m}(k\boldsymbol{r}) \equiv j_n(kr)\, Y_{n,m}(\theta,\phi)\ . \tag{6.154}$$

One can construct a row 'matrices' composed of the $\Psi_{\mathscr{J},n,m}(k\boldsymbol{r})$ or $\Psi_{\mathscr{H},n,m}(k\boldsymbol{r})$ functions respectively then the translation-addition theorem for scalar partial waves can be compactly expressed in matrix form:

$$\begin{aligned}
\Psi^t_{\mathscr{H}}(k\boldsymbol{r}) &= \Psi^t_{\mathscr{H}}\left(k\boldsymbol{r}'\right).\alpha(k\boldsymbol{r}_0) && r' > r_0\\
\Psi^t_{\mathscr{H}}(k\boldsymbol{r}) &= \Psi^t_{\mathscr{J}}\left(k\boldsymbol{r}'\right).\beta(k\boldsymbol{r}_0) && r' < r_0\\
\Psi^t_{\mathscr{J}}(k\boldsymbol{r}) &= \Psi^t_{\mathscr{J}}\left(k\boldsymbol{r}'\right).\beta(k\boldsymbol{r}_0) && \forall\left|k\boldsymbol{r}_j\right|\ ,
\end{aligned} \tag{6.155}$$

where the elements of the irregular translation-addition matrix, α have extremely simple expressions in terms of the $3Y$ coefficients :

$$\alpha_{\nu,\mu,nm}(k\boldsymbol{r}_0) = 4\pi i^{\nu-n} \sum_{p=|n-\nu|}^{n+\nu} i^p 3Y(n,m;\nu,\mu;p)\, h_p(kr_0)\, Y_{p,m-\mu}(\theta_0,\phi_0)\ , \tag{6.156}$$

while the elements of the regular translation-addition matrix, $\beta_{\nu mu,;n,m}$, are the same as the $\alpha_{\nu,\mu;n,m}$ coefficients but with the $j_p(kr_0)$ replacing the $h_p(kr_0)$ function i.e.:

$$\beta_{\nu,\mu;n,m}(k\boldsymbol{r}_0) \equiv 4\pi i^{\nu-n} \sum_{p=|n-\nu|}^{n+\nu} i^p 3Y(n,m;\nu,\mu;p)\, j_p(kr_0)\, Y_{p,m-\mu}(\theta_0,\phi_0)\ , \tag{6.157}$$

where $3Y\left(n,m;\nu,\mu;p\right)$ are the 3Y coefficients defined by the angular integral of three scalar spherical harmonics:

$$\begin{aligned} 3Y\left(n,m;\nu,\mu;p\right) &\equiv \int_0^{\pi}\int_0^{2\pi} Y_{n,m}\left(\theta,\phi\right) Y^*_{\nu,\mu}\left(\theta,\phi\right) Y^*_{p,m-\mu}\left(\theta,\phi\right)\sin\theta d\theta d\phi \\ &= (-)^{\mu}(-)^{m-\mu}\int_0^{\pi}\int_0^{2\pi} Y_{n,m}\left(\theta,\phi\right) Y_{\nu-\mu}\left(\theta,\phi\right) Y_{p,\mu-m}\left(\theta,\phi\right)\sin\theta d\theta d\phi \\ &= (-)^{m}\left[\frac{(2n+1)(2\nu+1)(2p+1)}{4\pi}\right]^{1/2}\begin{pmatrix} n & \nu & p \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} n & \nu & p \\ -m & \mu & \mu-m \end{pmatrix}. \end{aligned} \tag{6.158}$$

The symbol,

$$\begin{pmatrix} n & \nu & p \\ m & \mu & M \end{pmatrix}, \tag{6.159}$$

stands for the Wigner 3J coefficients. It is worth remarking that the '3Y' coefficients of eq.(6.158) are closely related to the Gaunt coefficients developed in quantum mechanics for treating the helium atom (mostly differing on account of different normalization conditions).

6.8.3 Vector translation-addition theorem

The vector translation-addition theorem are the vector analogue of the scalar translation theorem discussed above in section6.8.2. This would be an almost trivial extension of the scalar addition theorem if we were working with solutions of the vector Helmholtz equation in a Cartesian basis like those of eqs.(6.7) and (6.8). The additional complication is due to the fact we want to the vector translation-addition theorem to act on the purely transverse waves like those of eq.(6.9). Defining column matrices, $\mathbf{\Psi}$, composed of the transverse vector partial waves, the vector translation-addition theorem is written:

$$\begin{aligned} \mathbf{\Psi}^t_{\mathscr{H}}\left(k\boldsymbol{r}\right) &= \mathbf{\Psi}^t_{\mathscr{H}}\left(k\boldsymbol{r}'\right) J(k\boldsymbol{r}_0) && r' > r_0 \\ \mathbf{\Psi}^t_{\mathscr{H}}\left(k\boldsymbol{r}\right) &= \mathbf{\Psi}^t_{\mathscr{J}}\left(k\boldsymbol{r}'\right) H(k\boldsymbol{r}_0) && r' < r_0 \\ \mathbf{\Psi}^t_{\mathscr{J}}\left(k\boldsymbol{r}\right) &= \mathbf{\Psi}^t_{\mathscr{J}}\left(k\boldsymbol{r}'\right) J(k\boldsymbol{r}_0) && \forall |r_0| \ , \end{aligned} \tag{6.160}$$

where the matrix $J(k\boldsymbol{r}_0)$ matrix can be expressed in terms of spherical scalar $\beta(k\boldsymbol{r}_0)$ matrices of eq.(6.157) (expressed in terms of spherical Hankel functions) while the $H(k\boldsymbol{r}_0)$ matrices can be expressed in terms of the $\alpha(k\boldsymbol{r}_0)$ matrices of eq.(6.156).

Explicitly, the $H(k\boldsymbol{r}_0)$ matrix can be expressed:

$$H(k\boldsymbol{r}_0) = \begin{bmatrix} A_{\nu,\mu;n,m}\left(kr_0,\theta_0,\phi_0\right) & B_{\nu,\mu;n,m}\left(kr_0,\theta_0,\phi_0\right) \\ B_{\nu,\mu;n,m}\left(kr_0,\theta_0,\phi_0\right) & A_{\nu,\mu;n,m}\left(kr_0,\theta_0,\phi_0\right) \end{bmatrix}. \tag{6.161}$$

The vector coefficients $A_{\nu,\mu;n,m}$ are then calculated using :

$$\begin{aligned} A_{\nu,\mu;n,m} = \frac{1}{2}\sqrt{\frac{1}{\nu\left(\nu+1\right)n\left(n+1\right)}}\Big[&2\mu m\alpha_{\nu,\mu;n,m} + \\ &+\sqrt{(n-m)\left(n+m+1\right)}\sqrt{(\nu-\mu)\left(\nu+\mu+1\right)}\alpha_{\nu,\mu+1;n,m+1} \\ &+\sqrt{(n+m)\left(n-m+1\right)}\sqrt{(\nu+\mu)\left(\nu-\mu+1\right)}\alpha_{\nu,\mu-1;n,m-1}\Big] \ . \end{aligned} \tag{6.162}$$

When filling up a matrix, with the $A_{\nu\mu,nm}$ coefficient, we should fill them up column by column (calculate all the ν,μ) elements for a fixed n,m. Then, for each n,m, the $B_{\nu\mu,nm}$ coefficients can be calculated from the previous (i.e. $\nu-1$) scalar coefficients :

$$\begin{aligned} B_{\nu,\mu,n,m} = -i\frac{1}{2}\sqrt{\frac{2\nu+1}{2\nu-1}\frac{1}{\nu(\nu+1)n(n+1)}}\Big[&2m\sqrt{(\nu-\mu)(\nu+\mu)}\alpha_{\nu-1,\mu;n,m} \\ &+\sqrt{(n-m)(n+m+1)}\sqrt{(\nu-\mu)(\nu-\mu-1)}\alpha_{\nu-1,\mu+1;n,m+1} \\ &-\sqrt{(n+m)}\sqrt{(n-m+1)}\sqrt{(\nu+\mu)(\nu+\mu-1)}\alpha_{\nu-1,\mu-1;n,m-1}\Big] \ . \end{aligned} \tag{6.163}$$

6.8.4 Rotation matrices

Under rotation, each of the four blocks of a vector translation-addition matrix transform following the rotation matrix, $\mathscr{D}(\alpha,\beta,\gamma)$, which is expressed in terms of the 3 Euler angles, α, β, and γ. The $\mathscr{D}(\alpha,\beta,\gamma)$ matrix elements are described in detail in ref.[7], and are block diagonal in the orbital (multipole) 'quantum' number, n :

$$[\mathscr{D}(\alpha,\beta,\gamma)]_{\nu,\mu,n,m} = \delta_{n,\nu}\exp(i\mu\alpha)\, d^{(n)}_{\mu m}(\beta)\exp(im\gamma) \ . \tag{6.164}$$

The elements $d^{(n)}_{\mu m}$ are standard,[7] and the $d^{(n)}_{\mu,m}$ term in the rotation matrices can be expressed in terms of the Jacobi polynomials[7] :

$$d^{(n)}_{\mu m}(\beta) = \left[\frac{(n+\mu)!(n-\mu)!}{(n+m)!(n-m)}\right]^{1/2}\left(\cos\frac{\beta}{2}\right)^{m+\mu}\times\left(\sin\frac{\beta}{2}\right)^{m-\mu}P^{(\mu-m,m+\mu)}_{n-\mu}(\cos\beta) \ . \tag{6.165}$$

6.9 Recurrence relations for special functions

Partial wave descriptions are composed of products of spherical harmonic and spherical Bessel types special functions. For a numerical analysis, it is important to calculate these functions rapidly and accurately. Recurrence relations prove to be a good manner to achieve this goal. Multipole expansions must be truncated to a given order $n_{\max}$, which determines the strength of spatial field variations. Inspection of the translation-addition theorem formulas show us that we will need to evaluate spherical harmonic and spherical Hankel functions up to order $2n_{\max}$.

6.9.1 Recurrence relations for associated Legendre polynomials

The recurrence relations for the $\widetilde{P}^m_n$ polynomials are are initialized with:

$$\widetilde{P}^0_0(u) = \sqrt{\frac{1}{4\pi}} \ . \tag{6.166}$$

We can then calculate all the $\widetilde{P}^n_n$ up to $2n_{\max}$ with the recurrence relation:

$$\widetilde{P}^n_n(u) = -\sqrt{\frac{2n+1}{2n}}\widetilde{P}^{n-1}_{n-1}(u) \ . \tag{6.167}$$

The $\widetilde{P}_n^m$ with $m = n-1$ are calculated via:

$$\widetilde{P}_n^{n-1}(u) = u\sqrt{2n+1}\widetilde{P}_{n-1}^{n-1} \ . \tag{6.168}$$

All the remaining $\widetilde{P}_n^m$ with $m = 1,...,n-2$ can be calculated for each $n = 3,...,2n_{\max}$ using the relation :

$$\widetilde{P}_n^m(u) = \sqrt{\frac{(2n+1)}{(n^2-m^2)}}\left[\sqrt{(2n-1)}x\widetilde{P}_{n-1}^m - \sqrt{\frac{(n-1)^2-m^2}{(2n-3)}}\widetilde{P}_{n-2}^m\right] \ . \tag{6.169}$$

Although we obtain the $\widetilde{P}_n^0$ in the above scheme, it can sometimes prove useful to obtain the normalized Legendre Polynomials through the recurrence relation :

$$\widetilde{P}_n^0(u) = \frac{1}{n}\left(u\sqrt{4n^2-1}\right)\widetilde{P}_{n-1}^0(u) - (n-1)\sqrt{\frac{2n+1}{2n-3}}\widetilde{P}_{n-2}^0(u) \ . \tag{6.170}$$

The $\widetilde{P}_n^m$ with negative values of m are calculated using :

$$\widetilde{P}_n^{-m}(u) = (-)^m\widetilde{P}_n^m(u) \ . \tag{6.171}$$

6.9.2 Logarithmic Bessel functions

The spherical Bessel, Neumann and Hankel functions of the Ricatti form, are simply these functions multiplied by their argument. The advantage of this form is that they have better limit properties for small arguments. Their definitions are respectively:

$$\psi_n(z) \equiv zj_n(z), \qquad \chi_n(z) \equiv zy_n(z), \qquad \xi_n(z) \equiv zh_n(z) \ . \tag{6.172}$$

Multiplying the logarithmic derivatives of these functions by their argument defines the functions, $\varphi^{(1)}$, $\varphi^{(2)}$, $\varphi^{(3)}$:

$$\varphi_n^{(1)}(z) \equiv \frac{\psi_n'(z)}{j_n(z)}, \qquad \varphi_n^{(2)}(z) \equiv \frac{\chi_n'(z)}{y_n(z)}, \qquad \varphi_n^{(3)}(z) \equiv \frac{\xi_n'(z)}{h_n(z)} \ . \tag{6.173}$$

The $\varphi_n^{(i)}$ can be generated by particularly efficient numerical recursion relations. They also provide particularly symmetric expression for the Mie coefficients of spherical scatterers and formulations of matrix balancing. The Wronskian relation for Ricatti-Bessel functions:

$$\begin{aligned}\psi_n(x)\,\xi_n'(x) - \psi_n'(x)\,\xi_n(x) &= i \\ \psi_n(x)\,\chi_n'(x) - \psi_n'(x)\,\chi_n(x) &= 1 \ ,\end{aligned} \tag{6.174}$$

takes a simple form in terms of the $\varphi_n^{(1,2,3)}$ functions as:

$$\begin{aligned}\varphi_n^{(3)}(z) - \varphi_n^{(1)}(z) &= \frac{i}{xj_n(x)\,h_n(x)} \\ \varphi_n^{(2)}(z) - \varphi_n^{(1)}(z) &= \frac{1}{xj_n(x)\,y_n(x)}\end{aligned} \tag{6.175}$$

The $\varphi^{(3)}$ can be relatively well be calculated numerically from the upward recurrence relation:

$$\varphi_n^{(3)}(z) = \frac{z^2}{n - \varphi_{n-1}^{(3)}(z)} - n \ , \tag{6.176}$$

starting with an initialization of

$$\varphi_0^{(3)}(z) = iz \ . \tag{6.177}$$

The $\xi_n(z)$ functions in most situations can then be readily calculated numerically from the recurrence relation:

$$\xi_n(z) = \frac{\xi_{n-1}(z)}{z}\left(n - \varphi_{n-1}^{(3)}(z)\right) \ , \tag{6.178}$$

starting from the initial value $\xi_0(z) = -ie^{iz}$. One should note that analytical expressions exist for the spherical Ricatti-Hankel functions, $\xi_n(z)$, and these can be useful at low multipole order:

$$\begin{aligned}
\xi_0(z) &= -ie^{iz} \\
\xi_1(z) &= -e^{iz}\left(1 + \frac{i}{z}\right) \\
\xi_2(z) &= e^{iz}\left(i - \frac{3}{z} - \frac{3i}{z^2}\right) \ .
\end{aligned} \tag{6.179}$$

When one deals with high multipole orders however, it usually is more practical to exploit the recurrence relations of eq.(6.176) and (6.178).

The regular Ricatti Bessel functions, $\varphi_n^{(1)}(z)$, obey the same recurrence relations as the $\varphi_n^{(3)}(z)$ functions. If one calculates them by via upward recurrence, things may work fine for the first few recurrence calculations, but at some point, the recurrence relations can go completely off course. The solution to this problem has been known for quite some time is that the $\varphi_n(z)$ functions should be calculated starting from high values of n in a reverse recurrence relation. I start usually with n equal to at least $n_{\max} + 20$ where $n_{\max}$ is the largest value that I want to use in calculations, and start simply with $\varphi_{n_{\max}+20}(z) = 0$. The $\varphi_n(z)$ functions so obtained have always been the correct ones up to machine precision. The reverse recurrence relation is:

$$\varphi_n(z) = n + 1 - \frac{z^2}{n + 1 + \varphi_{n+1}(z)} \ . \tag{6.180}$$

One can check calculations by verifying that the $\varphi_0(z)$ obtained by backward recurrence is equal to the analytical result:

$$\varphi_0(z) = z\frac{\cos z}{\sin z} \ . \tag{6.181}$$

Once the $\varphi_n(z)$ functions have been calculated, one can readily generate the $\psi_n(z)$ functions with the upward recurrence relation which is the direct analogue of eq.(6.178)

$$\psi_n(z) = \frac{\psi_{n-1}(z)}{z}\left(n - \varphi_{n-1}(z)\right) \ , \tag{6.182}$$

starting with the initial value $\psi_0(z) = \sin z$. Analytical expressions for the lowest $\psi_n(z)$ are:

$$\begin{aligned}
\psi_0(z) &= \sin z \\
\psi_1(z) &= \frac{\sin z}{z} - \cos z \qquad (6.183) \\
\psi_2(z) &= \left(\frac{3}{z^2} - 1\right)\sin z - \frac{3}{z}\cos z\,. \qquad (6.184)
\end{aligned}$$

It is perhaps worth remarking that, there is one potential numerical problem with using eq.(6.178) to calculate spherical Hankel functions. For some values of z the real and imaginary parts of the spherical Hankel functions can have extremely different absolute values. For concreteness, let us assume that $|\mathrm{Im}(h_n(z))| \ll |\mathrm{Re}(h_n(z))|$, then using eq.(6.178), the calculated value of $\mathrm{Im}(h_n(z))$ will usually be quite inaccurate if its absolute value is less than last significant figure in the calculation of $\mathrm{Re}(h_n(z))$. This problem can be circumvented (for real values of z at least) by calculating the spherical Neumann functions (denoted here by $y_n(z)$ but some authors denote it $n_n(z)$). We recall that the Neumann functions are real-valued provided that z is real valued.

The Ricatti Neumann functions are defined :

$$\chi_n(z) \equiv z y_n(z)\ . \qquad (6.185)$$

The first few Ricatti Neumann functions, $\chi_n(z)$, are

$$\begin{aligned}
\chi_0(z) &= -\cos z \\
\chi_1(z) &= -\frac{\cos z}{z} - \sin z \\
\chi_2(z) &= -\left(\frac{3}{z^2} - 1\right)\cos z - \frac{3}{z}\sin z\,. \qquad (6.186)
\end{aligned}$$

We define a $\varphi^{(2)}$ 'logarithmic derivative' Neumann function as :

$$\varphi_n^{(2)}(z) \equiv \frac{\chi_n'(z)}{y_n(z)}\ . \qquad (6.187)$$

We can calculate the $\varphi_n^{(2)}$ from the upward recurrence relation:

$$\varphi_n^{(2)}(z) = \frac{z^2}{n - \varphi_{n-1}^{(2)}(z)} - n\ , \qquad (6.188)$$

with an initialization of

$$\varphi_0^{(2)}(z) = -z\frac{\sin z}{\cos z}\ . \qquad (6.189)$$

Once the $\varphi_n^{(2)}$ functions have been calculated, one can readily generate the $\chi_n(z)$ functions with the upward recurrence relation which is the direct analogue of eq.(6.178)

$$\chi_n(z) = \frac{\chi_{n-1}(z)}{z}\left(n - \varphi_{n-1}^{(2)}(z)\right)\ , \qquad (6.190)$$

starting with the initial value $\chi_0(z) = -\cos z$.

Having calculated $\psi_n(z)$ via eq.(6.182) and $\chi_n(z)$ via eq.(6.190), one can finally construct $h_n(z)$ by

$$h_n(z) \equiv j_n(z) + iy_n(z) \ , \tag{6.191}$$

with the real and imaginary parts of $\xi_n(z)$ now both being calculated up to machine accuracy.

The ratio of spherical Bessel functions to spherical Hankel functions also occurs frequently in Mie theory, and I found it convenient and more accurate to calculate these ratios directly using upward recurrence relations, notably:

$$\frac{j_n(z)}{h_n(z)} = \frac{j_{n-1}(z)}{h_{n-1}(z)} \frac{n - \varphi_{n-1}(z)}{n - \varphi_{n-1}^{(3)}(z)} \ , \tag{6.192}$$

with the initialization

$$\frac{j_0(z)}{h_0(z)} = i \sin z \exp[-iz] = \frac{1 - e^{-2iz}}{2} \tag{6.193}$$

$$\frac{j_1(z)}{h_1(z)} = \frac{1}{2}\left(1 - \frac{i-z}{i+z} e^{-2iz}\right) \ , \tag{6.194}$$

which has good properties for numerical calculations. For example, the expression and $j_1(z)/h_1(z)$ satisfies the recurrence relation if we start with $j_0(z)/h_0(z)$ since

$$\begin{aligned} \frac{j_1(z)}{h_1(z)} &= \frac{j_0(z)}{h_0(z)} \frac{1 - \varphi_0(z)}{1 - \varphi_0^{(3)}(z)} = \frac{1 - e^{-2iz}}{2} \frac{1 - z\frac{\cos z}{\sin z}}{1 - iz} = -\frac{1}{2i} \frac{1 - iz - e^{-2iz} - ize^{-2iz}}{i + z} \\ &= \frac{1}{2}\left(1 - \frac{i-z}{i+z} e^{-2iz}\right) . \end{aligned} \tag{6.195}$$

For coated spheres, it is can also useful to use the analogous recurrence relation:

$$\frac{j_n(z)}{y_n(z)} = \frac{\psi_{n-1}(z)}{y_{n-1}(z)} \frac{n - \varphi_{n-1}(z)}{n - \varphi_{n-1}^{(2)}(z)} \ , \tag{6.196}$$

with the initialization of

$$\frac{j_0(z)}{y_0(z)} = -\tan z \ , \tag{6.197}$$

with the first recurrence giving

$$\frac{j_1(z)}{y_1(z)} = \frac{z\cos z - \sin z}{z\sin z + \cos z} \ . \tag{6.198}$$

Other useful relations are obtained from the classic spherical Bessel function recurrence relations:

$$\begin{aligned} f_n(z) &= \frac{z f_{n-1}(z) + z f_{n+1}(z)}{2n+1} \\ f_n'(z) &= \frac{n f_{n-1}(z) - (n+1) f_{n+1}(z)}{2n+1} \ , \end{aligned} \tag{6.199}$$

with $f_n(z) = j_n(z), h_n(z)$ to obtain :

$$z j_n'(z) + (n+1) j_n(z) = n\frac{\psi_{n-1}(z)}{2n+1} + (n+1)\frac{\psi_{n-1}(z)}{2n+1} = \psi_{n-1}(z) \ , \tag{6.200}$$

with

$$\psi_n'(z) \equiv z j_n'(z) + j_n(z) \tag{6.201}$$

we obtain a convenient expression for the derivative of Ricatti-Bessel functions:

$$\psi_n'(z) = \psi_{n-1}(z) - n j_n(z) \ . \tag{6.202}$$

or the expression:

$$\psi_n'(z) = (n+1) j_n(z) - \psi_{n+1}(z) \ . \tag{6.203}$$

Since the recurrence relations of eq.(6.182) and (6.190) are numerically stable, these relations give us a convenient way to calculate $\varphi_n^{(3)}(z)$:

$$\varphi_0^{(3)}(z) = iz \qquad\qquad \varphi_n^{(3)}(z) = \frac{\psi_{n-1}(z) + i y_{n-1}(z)}{\psi_n(z) + i y_n(z)} - n$$

6.9.3 Vector Spherical Harmonics

There is no universally accepted notation for the Vector Spherical Harmonics (VSHs). Our notation for their *normalized* forms is $\boldsymbol{X}_{n,m}$, $\boldsymbol{Y}_{n,m}$, and $\boldsymbol{Z}_{n,m}$ where they are respectively defined by

$$\begin{aligned}
\boldsymbol{X}_{n,m}(\theta,\phi) &\equiv \boldsymbol{Z}_{n,m}(\theta,\phi) \times \widehat{\boldsymbol{r}} \\
\boldsymbol{Y}_{n,m}(\theta,\phi) &\equiv \widehat{\boldsymbol{r}} Y_{n,m}(\theta,\phi) \\
\boldsymbol{Z}_{n,m}(\theta,\phi) &\equiv \frac{r\boldsymbol{\nabla} Y_{n,m}(\theta,\phi)}{\sqrt{n(n+1)}} = \widehat{\boldsymbol{r}} \times \boldsymbol{X}_{n,m}(\theta,\phi) \ ,
\end{aligned} \tag{6.204}$$

The scalar spherical harmonics, $Y_{n,m}(\theta,\phi)$, do have a nearly universal convention for their definitions[10] which we recalled in eq.(6.138).

For numerical calculations of the VSHs, it is convenient to introduce the *normalized* functions $\overline{u}_n^m$ and $\overline{s}_n^m$ defined as:

$$\overline{u}_n^m(\cos\theta) \equiv \gamma_{n,m}\frac{m}{\sin\theta}P_n^m(\cos\theta) \tag{6.205}$$

$$\overline{s}_n^m(\cos\theta) \equiv \gamma_{n,m}\frac{d}{d\theta}P_n^m(\cos\theta) \ , \tag{6.206}$$

where the P_n^m are the Legendre functions defined in eqs.(6.141), (6.142), and (6.143), and $\gamma_{n,m}$ a normalization factor given by

$$\gamma_{n,m} \equiv \frac{\lambda_{n,m}}{\sqrt{n(n+1)}} = \sqrt{\frac{(2n+1)(n-m)!}{4\pi n(n+1)(n+m)!}} \ . \tag{6.207}$$

The transverse VSHs have compact expressions in terms of $\overline{u}_n^m$ and $\overline{s}_n^m$:

$$\begin{aligned}\boldsymbol{X}_{n,m}(\theta,\phi) &= \left[i\overline{u}_n^m(\cos\theta)\,\widehat{\boldsymbol{\theta}} - \overline{s}_n^m(\cos\theta)\widehat{\boldsymbol{\phi}}\right]\exp(im\phi)\\ \boldsymbol{Z}_{n,m}(\theta,\phi) &= \left[\overline{s}_n^m(\cos\theta)\widehat{\boldsymbol{\theta}} + i\overline{u}_n^m(\cos\theta)\,\widehat{\boldsymbol{\phi}}\right]\exp(im\phi)\ .\end{aligned} \tag{6.208}$$

The normalized $\overline{u}_n^m$ functions can be readily calculated with recurrence relations:

$$\begin{aligned}\overline{u}_n^0(x) &= 0\ , \qquad \overline{u}_1^1(x) = -\frac{1}{4}\sqrt{\frac{3}{\pi}}\\ \overline{u}_n^n(x) &= -\sqrt{\frac{n\,(2n+1)}{2\,(n+1)\,(n-1)}}\sqrt{1-x^2}\,\overline{u}_{n-1}^{n-1}(x)\\ \overline{u}_n^m(x) &= \sqrt{\frac{(n-1)\,(4n^2-1)}{(n+1)\,(n^2-m^2)}}x\,\overline{u}_{n-1}^m(x)\\ &\qquad -\sqrt{\frac{(2n+1)\,(n-1)\,(n-2)\,(n-m-1)\,(n+m-1)}{(2n-3)\,n\,(n+1)\,(n^2-m^2)}}\overline{u}_{n-2}^m(x)\\ \overline{u}_n^{n-1}(x) &= \sqrt{\frac{(2n+1)\,(n-1)}{(n+1)}}x\overline{u}_{n-1}^{n-1}(x)\ ,\end{aligned} \tag{6.209}$$

while the $\overline{s}_n^m$ can be determined from the $\overline{u}_n^m$ functions via the formula:

$$\overline{s}_n^m(\cos\theta) = \frac{1}{(m+1)}\sqrt{(n+m+1)\,(n-m)}\sin\theta\,\overline{u}_n^{m+1}(\cos\theta) + \cos\theta\,\overline{u}_n^m(\cos\theta)\ . \tag{6.210}$$

The respective parity properties of $\overline{u}_n^m$ and $\overline{s}_n^m$ are:

$$\begin{aligned}\overline{u}_n^{-m}(x) &= (-1)^{m+1}\,\overline{u}_n^m(x)\\ \overline{s}_n^{-m}(x) &= (-1)^m\,\overline{s}_n^m(x)\ .\end{aligned} \tag{6.211}$$

References:

[1] I.A. Abramowitz, M.; Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications, 1972.

[2] A. Archambault, T. V. Teperik, F. Marquier, and J. J. Greffet. Surface plasmon fourier optics. *Phys. Rev. B*, 79:195414, May 2009.

[3] J. M. Borwein, M. L. Glasser, R. C. McPhedran, J. G. Wan, and I. J. Zucker. *Lattice Sums : Then and Now*. Series Encyclopedia of Mathematics and its Applications (No. 150), 2013.

[4] Salvatore Campione, Sergiy Steshenko, and Filippo Capolino. Complex bound and leaky modes in chains of plasmonic nanospheres. *Opt. Express*, 19(19):18345–18363, Sep 2011.

[5] Weng Cho Chew. *Waves and Fields in Inhomogeneous Media*. IEEE Press, New York, 1990.

[6] Matteo Conforti and Massimiliano Guasoni. Dispersive properties of linear chains of lossy metal nanoparticles. *J. Opt. Soc. Am. B*, 27:1576–1582, 2010.

[7] A. R. Edmonds. *Angular Momentum in Quantum Mechanics*. Princeton University Press, Princeton New Jersey, 1960.

[8] S. Enoch, R. C. McPhedran, N.A. Nicorovici, L. C. Botten, and J. N. Nixon. Sums of spherical waves for lattices, layers and lines. *J. Math. Phys.*, 42:5859–5870, 2001.

[9] Ana L. Fructos, Salvatore Campione, Filippo Capolino, and Francisco Mesa. Characterization of complex plasmonic modes in two-dimensional periodic arrays of metal nanospheres. *J. Opt. Soc. Am. B*, 28(6):1446–1458, Jun 2011.

[10] J. D. Jackson. *Classical Electrodynamics : Third Edition*. John Wiley & Sons, 1999.

[11] A. Femius Koenderink and Albert Polman. Complex response and polariton-like dispersion splitting in periodic metal nanoparticle chains. *Phys. Rev. B*, 74:033402(4), 2006.

[12] A.F. Koenderink. Plasmon nanoparticle array waveguides for single photon and single plasmon sources. *Nano Lett.*, 9(12):4228–4233, 2009.

[13] W. Kohn and N. Rostoker. Solution of the schrödinger equation in periodic lattices with an application to metallic lithium. *Phys. Rev.*, 94:1111–1120, Jun 1954.

[14] J Korringa. On the calculation of the energy of a bloch wave in a metal. *Physica*, 13(67):392 – 400, 1947.

[15] C. M. Linton. Lattice sums for the helmholtz equation. *Society for Industrial and Applied Mathematics*, 52:630–674, 2010.

[16] C. M. Linton and I. Thompson. One- and two-dimensional lattice sums for the three-dimensional helmholtz equation. *J. Comput. Physics*, 228:1815–1829, 2009.

[17] M.I. Mischenko, G. Videen, V. A. Babenko, N. G. Khlebtsov, and T. Wriedt. T-matrix theory of electromagnetic scattering by particles and its applications: a comprehensive reference database. *J. Quant. Spect. Rad. Trans.*, 88:357–406, 2004.

[18] M. I. Mishchenko, L. D. Travis, and D. W. Mackowski. T-matrix computations of light scattering by nonspherical particles: A review. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 55:535 – 575, 1996.

[19] Alexander Moroz. Exponentially convergent lattice sums. *Opt. Lett.*, 26(15):1119–1121, Aug 2001.

[20] Alexander Moroz. Quasi-periodic green's functions of the helmholtz and laplace equations. *Journal of Physics A: Mathematical and General*, 39(36):11247, 2006.

[21] R.G. Newton. *Scattering Theory of Waves and Particles*. McGraw-Hill New York, 1966.

[22] N Papanikolaou, R Zeller, and P H Dederichs. Conceptual improvements of the kkr method. *Journal of Physics: Condensed Matter*, 14(11):2799, 2002.

[23] Brice Rolly, Nicolas Bonod, and Brian Stout. Dispersion relations in metal nanoparticle chains: necessity of the multipole approach. *J. Opt. Soc. Am. B*, 29(5):1012–1019, May 2012.

[24] Ping Sheng. *Introduction to Wave Scattering, Localization and Mesoscopic Phenomena*. Springer, 2005.

[25] J. C. Slater. Wave functions in a periodic potential. *Physical Review*, 51:846–851, 1937.

[26] N Stefanou and A Modinos. Scattering of light from a two-dimensional array of spherical particles on a substrate. *Journal of Physics: Condensed Matter*, 3(41):8135, 1991.

[27] N Stefanou and A Modinos. Scattering of electromagnetic waves by a disordered two-dimensional array of spheres. *Journal of Physics: Condensed Matter*, 5(47):8859, 1993.

[28] N. Stefanou, V. Yannopapas, and A. Modinos. Heterostructures of photonic crystals: frequency bands and transmission coefficients. *Computer Physics Communications*, 113(1):49 – 77, 1998.

[29] B. Stout, J.-C. Auger, and J. Lafait. A transfer matrix approach to local field calculations in multiple scattering problems. *J. Mod. Opt.*, 49:2129–2152, 2002.

[30] B. Stout, J.C. Auger, and A. Devilez. Recursive t-matrix algorithm for resonant multiple scattering: Applications to localized plasmon excitations. *J. Opt. Soc. Am. A*, 25:2549–2557, 2008.

[31] L. Tsang, J. A. Kong, and R. T. Shin. *Theory of Microwave Remote Sensing*. John Wiley & Sons New York, 1985.

[32] Serge Winitzki. *Computing the incomplete gamma function to arbitrary precision*. Springer, Berlin, Lecture Notes in Comput. Sci. 2667, 2003.

[33] R.C. Wittmann. Spherical wave operators and the translation formulas. *Antennas and Propagation, IEEE Transactions on*, 36(8):1078 –1087, aug 1988.

Chapter 7:
Differential Method for Periodic Structures

Evgeny Popov

Table of Contents:

Chapter 7

Differential Method for Periodic Structures

Evgeny Popov

Aix-Marseille Université, CNRS Central Marseille, Institut Fresnel UMR 7249
Campus de Saint Jerome,13013 Marseille, France
e.popov@fresnel.fr *www.fresnel.fr/perso/popov*

The basic idea of the differential methods consists in projecting the electromagnetic field on a set of basic functions in order to reduce Maxwell partial differential equations into a set of ordinary differential equations. When working in a Cartesian coordinates, the natural basis consists of exponentials, using the periodicity of the optogeometrical parameters. Diffraction by a single aperture requires working in the basis of cylindrical Bessel functions [7.1], while diffraction by an arbitrary-shaped single object requires vector spherical functions [7.2] as a basis.

The first studies using the differential method [7.3] appeared in the late 1960s, initiated by the birth of the computers. These studies concerned the modeling of diffusion of particles in nuclear potential by using the separation of variables of the radial Schrödinger equation. The method was called "optical method" due to the similarity between the Schrödinger and the Helmholtz equations. The first applications to grating diffraction appear in 1969 [7.4], but accurate and converging results required combining the differential method with conformal mapping techniques [7.5]. The classical differential theory as known nowadays was formulated in [7.6, 7.7]. One can find a detailed review on the classical differential method in [7.8]

It appeared that the classical differential theory suffered from severe numerical problems in transverse magnetic (TM) polarization, as well as for deep gratings. The first breakthrough was made in the first half of the 1990s, by introducing orthonormalization of the differential equations during their integration [7.9] and followed later by the so-called R-matrix or S-matrix propagating algorithms [7.10]. The second breakthrough improved considerably the convergence in TM polarization for lamellar gratings, by introducing the correct factorization rules (see further on), at first by chance [7.11] and after that using theoretical arguments [7.12], closely followed by a generalization to arbitrary profiles [7.13]. A detailed review can be found in [7.14].

7.1. Maxwell equations in the truncated Fourier space

Let us consider a structure with two-dimensional periodicity along the x- and y-axis (Fig.7.1) with periods equal to d_x and d_y. The modulated (grating) region extends in z from z_{min} to z_{max}. Inside that region, for a given value of the vertical coordinate z, the permittivity ε and permeability μ are periodic functions in x and y that can be projected on exponential Fourier basis:

$$\varepsilon(x,y,z)=\sum_{m,n=-\infty}^{+\infty}\varepsilon_{m,n}(z)\exp(imK_x x+inK_y y)$$
$$\mu(x,y,z)=\sum_{m,n=-\infty}^{+\infty}\mu_{m,n}(z)\exp(imK_x x+inK_y y) \qquad (7.1)$$

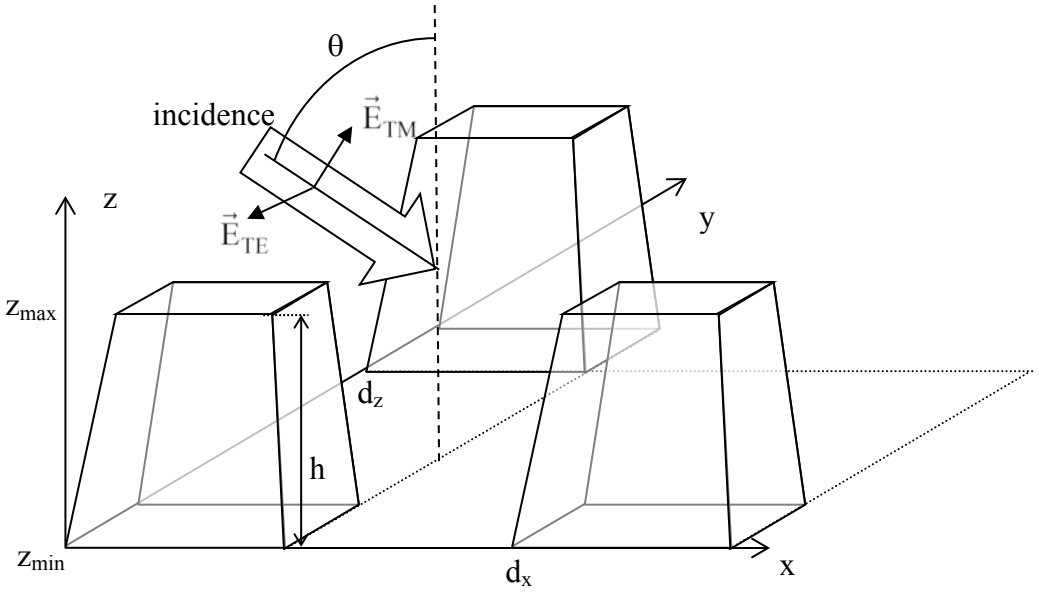


Fig.7.1. Schematical representation of a structure having two-dimensional periodicity in x and y-directions, consisting of truncated pyramids with height h.

where $K_x=2\pi/d_x$ and $K_y=2\pi/d_y$. We shall deal with a monochromatic (wavelength λ) plane wave incident on the structure with a wavevector:

$$\vec{k}_{inc}=(\alpha_0,\beta_0,-\gamma_0) \qquad (7.2)$$

with components related to the incident polar angle θ (between the incident direction and the grating normal) and azimuthal angle φ (between the plane of incidence and the xOz-plane):

$$\alpha_0=k_0\sin\theta\cos\varphi,\quad \beta_0=k_0\sin\theta\cos\varphi,$$
$$\gamma_0=\sqrt{k_0^2n_{inc}^2-\alpha_0^2-\beta_0^2},\quad k_0=2\pi/\lambda \qquad (7.3)$$

where n_{inc} is the refractive index of the cladding.

The existence and uniqueness of the solution of the diffraction problem is an interesting problem that is not discussed here. The reader can refer to several basic works (see for example [7.15, 7.16]. What is important to conclude is that the electromagnetic field is pseudo-periodic, so that similarly to eq.(7.1), the electric $\vec{E}$ and magnetic $\vec{H}$ field vectors can be represented in pseudo-Fourier series:

$$\vec{E}(x,y,z)=\sum_{m,n=-\infty}^{+\infty}\vec{E}_{m,n}(z)\exp\left[i(\alpha_0+mK_x)x+i(\beta_0+nK_y)y\right]$$
$$\vec{H}(x,y,z)=\sum_{m,n=-\infty}^{+\infty}\vec{H}_{m,n}(z)\exp\left[i(\alpha_0+mK_x)x+i(\beta_0+nK_y)y\right] \tag{7.4}$$

In what follows, we use the notations:

$$\alpha_m=\alpha_0+mK_x,\quad \beta_n=\beta_0+nK_y. \tag{7.5}$$

From a numerical point of view, it is necessary to truncate the series in eqs.(7.1) and (7.4), introducing truncation parameters N_x and N_y, which limit the lower and the upper boundaries in the series.

Maxwell equations written in Fourier space take the form, assuming $\exp(-i\omega t)$ time dependence with circular frequency ω:

$$\begin{aligned}
&i\beta_{m,n}E_{z,m,n}(z)-\frac{d}{dz}E_{y,m,n}(z)=i\omega B_{x,m,n}(z)\\
&\frac{d}{dz}E_{x,m,n}(z)-i\alpha_{m,n}E_{z,m,n}(z)=i\omega B_{y,m,n}(z)\\
&i\alpha_{m,n}E_{y,m,n}(z)-i\beta_{m,n}E_{x,m,n}(z)=i\omega B_{z,m,n}(z)\\
&i\beta_{m,n}H_{z,m,n}(z)-\frac{d}{dz}H_{y,m,n}(z)=-i\omega D_{x,m,n}(z)\\
&\frac{d}{dz}H_{x,m,n}(z)-i\alpha_{m,n}H_{z,m,n}(z)=-i\omega D_{y,m,n}(z)\\
&i\alpha_{m,n}H_{y,m,n}(z)-i\beta_{m,n}H_{x,m,n}(z)=-i\omega D_{z,m,n}(z).
\end{aligned} \tag{7.6}$$

As can be observed, the third and the sixth equations are not differential equations, and they are used to eliminate the z-components of the fields, as shown further on. It has to be stressed out that the equations with different (m,n) numbers are coupled through the Fourier components of $\vec{D}=\varepsilon\vec{E}$ and $\vec{B}=\mu\vec{H}$.

The next step is to factorize the products $\vec{D}=\varepsilon\vec{E}$ and $\vec{B}=\mu\vec{H}$. In this chapter we assume media with linear dielectric and magnetic properties and without spontaneous polarizations. The problem of Fourier transform of the product of two functions

$$\vec{D}(x,y,z)=\sum_{m,n=-\infty}^{+\infty}\vec{D}_{m,n}(z)\exp\left[i(\alpha_0+mK_x)x+i(\beta_0+nK_y)y\right] \tag{7.7}$$

is, in generally, solved theoretically by convolution of the Fourier transformers of the two functions, using the so-called Laurent's rule:

$$\vec{D}_{m,n}(z)=\sum_{m',n'=-\infty}^{+\infty}\varepsilon_{m-m',n-n'}(z)\vec{E}_{m',n'}(z). \tag{7.8}$$

However, there are several problems in the numerical application of this rule:

First, numerical applications are simplified when using matrix notations. However, most of the standard routines use single-rank vectors and rectangular (2-ranks) matrices, while the vectors D and E in eq.(7.8) have two indexes, and the matrix ε depend on four indexes. In the case of classical grating with one-dimensional periodicity, this problem does not exist. Fortunately, for structures having 2D periodicity, a reduction to standard arrays is possible by introduction of a single index instead of the double for the vectors, by the following substitution:

$$p = (m + N_x)(2N_y + 1) + (n + N_y + 1) \tag{7.9}$$

so that when m varies between $-N_x$ and $+N_x$ and n varies between $-N_y$ and $+N_y$, p varies between 1 and $P_{max} = (2N_x+1)(2N_y+1)$. Using these notations, we can introduce standard arrays in the following manner:

$$\begin{aligned} &\vec{D}_p(z) = \vec{D}_{m,n}(z), \quad \vec{E}_p(z) = \vec{E}_{m,n}(z), \quad \text{etc. for } \vec{H} \text{ and } \vec{B}, \\ &\varepsilon_{p-p'}(z) = \varepsilon_{m-m',n-n'}(z) \end{aligned} \tag{7.10}$$

so that eq.(7.8) takes the standard truncated form

$$\vec{D}_p(z) = \sum_{p'=1}^{P_{max}} \varepsilon_{p-p'}(z)\vec{E}_{p'}(z)\,. \tag{7.11}$$

That can be written in matrix notations in the form

$$\left[\vec{D}(z)\right] = \left[\!\left[\varepsilon(z)\right]\!\right]\left[\vec{E}(z)\right], \tag{7.12}$$

where double square brackets stand for the Toepliz matrix.

In addition, two diagonal matrices are useful:

$$\begin{aligned} \alpha_{p,p'} &= \delta_{p,p'}\alpha_m \\ \beta_{p,p'} &= \delta_{p,p'}\beta_n \end{aligned}, \tag{7.13}$$

with $\delta_{p,p'}$ being the Kronecker's symbol.

Second, due to the vectorial character of the fields, the matrix form in eq.(7.12) has to be interpreted in a block form:

$$\begin{pmatrix} [D_x(z)] \\ [D_y(z)] \\ [D_z(z)] \end{pmatrix} = \left[\!\left[\varepsilon(z)\right]\!\right] \begin{pmatrix} [E_x(z)] \\ [E_y(z)] \\ [E_z(z)] \end{pmatrix}, \text{ isotropic media} \tag{7.14}$$

$$\begin{pmatrix} [D_x(z)] \\ [D_y(z)] \\ [D_z(z)] \end{pmatrix} = \begin{pmatrix} \left[\!\left[\varepsilon_{xx}(z)\right]\!\right] & \left[\!\left[\varepsilon_{xy}(z)\right]\!\right] & \left[\!\left[\varepsilon_{xz}(z)\right]\!\right] \\ \left[\!\left[\varepsilon_{yx}(z)\right]\!\right] & \left[\!\left[\varepsilon_{yy}(z)\right]\!\right] & \left[\!\left[\varepsilon_{yz}(z)\right]\!\right] \\ \left[\!\left[\varepsilon_{zx}(z)\right]\!\right] & \left[\!\left[\varepsilon_{zy}(z)\right]\!\right] & \left[\!\left[\varepsilon_{zz}(z)\right]\!\right] \end{pmatrix} \begin{pmatrix} [E_x(z)] \\ [E_y(z)] \\ [E_z(z)] \end{pmatrix}, \text{ anisotropic media.} \tag{7.15}$$

The **third** problem linked with the truncation of eq. (7.8) has limited the use of the differential methods (including RCW method) for more than 30 years, and has been solved for lamellar gratings in the late 90s [7.11, 7.12], and for arbitrary-profile gratings in the start of the 2000s [7.13]. The problem is due to the very slow convergence with respect to the number of Fourier components in the truncated sum of eq. (7.8), when the two functions in the product are discontinuous. As demonstrated by Li [7.12], four different cases can be distinguished with respect to eq.(7.12):

1. Both ε and E are continuous functions of x and y.
2. ε is discontinuous, but E is continuous. This is the case of the tangential component of E.
3. Both ε and E are discontinuous, but their product D is continuous, as it happens for the normal component of D.
4. All three functions are discontinuous.

In the first and second case, Laurent's rule assures relatively rapid convergence. In the third case, more rapidly converging scheme can be obtained through the following considerations for isotropic media.

If D = εE is continuous, then it is possible to factorize the product between D (continuous) and 1/ε (discontinuous) using the Laurent's rule (called by Li *direct* rule):

$$\left[\vec{E}(z)\right] = \left[\!\left[\frac{1}{\varepsilon(z)}\right]\!\right]\left[\vec{D}(z)\right], \tag{7.16}$$

wherefrom the so called *inverse* rule is formulated:

$$\left[\vec{D}(z)\right] = \left[\!\left[\frac{1}{\varepsilon(z)}\right]\!\right]^{-1}\left[\vec{E}(z)\right], \tag{7.17}$$

which can be applied if the matrix $\left[\!\left[\frac{1}{\varepsilon(z)}\right]\!\right]$ is not singular, a requirement that can create numerical problem for highly conducting gratings having small imaginary part of ε.

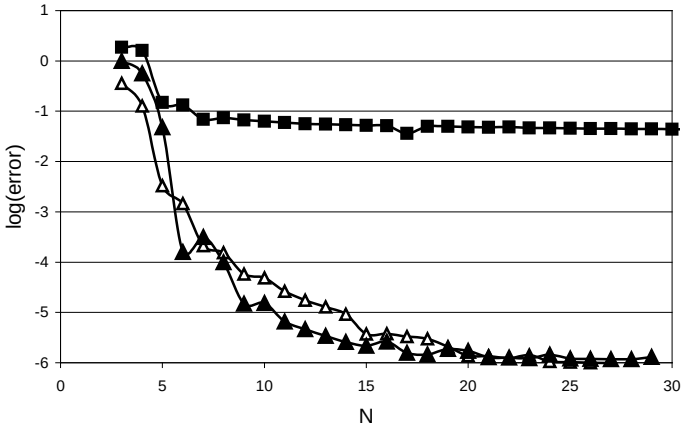


Fig.7.2.Convergence of the classical and the FFF version of the differential theory in the case of a dielectric sinusoidal grating with high contrast. Squares, old version of the differential theory for TM polarization; open triangles, new version, TM polarization; solid triangles, TE polarization (after [7.13]).

When infinite series are considered, eq.(7.17) is identical with eq.(7.12). However, as shown in Fig.7.2, the correct use of the direct or the inverse rules improves drastically the convergence of the differential methods with respect to the truncation parameter. Similarly to the abbreviation FFT, standing for Fast Fourier transformation, we have introduced the term Fast Fourier factorization (FFF) to name the correct use of the direct and the inverse rules, when applied numerically in the truncated Fourier space.

In the fourth case, neither the direct, nor the inverse rule result in acceptable convergence, so that this case must be avoided. Fortunately, this can be done by considering separately the electromagnetic field components, tangential or normal to the grating profile and taking into account that the electric field components tangential to the surface separating two different permittivities are continuous, in the same way as the normal components of the displacement D.

7.2. Differential theory for crossed gratings made of isotropic materials

In the isotropic case, the displacement vector $\vec{D}$ can easily be separated into a continuous part $\vec{D}_N = \varepsilon\vec{E}_N$, normal to the profile surface, and $\vec{D}_T = \varepsilon\vec{E}_T$ that contains the continuous function $\vec{E}_T$. Let us define a unit vector $\vec{N}$, normal to the grating profile. Although it is well defined on the profile (except edges), it is necessary to generalize its definition all over the grating region, which cannot be done in a unique manner. Different choices are shown further on for specific gratings having 1D or 2D periodicity. Using this generalized vector, the relations between $\vec{E}$ and $\vec{D}$ can be decomposed into two terms, for each of which we can apply the direct or the inverse factorization rules, skipping the explicit writing of the z-dependence:

$$\vec{D} = \varepsilon\vec{E}_N + \varepsilon\vec{E}_T = \varepsilon\vec{N}\left(\vec{N}\cdot\vec{E}\right) + \varepsilon\left[\vec{E} - \vec{N}\left(\vec{N}\cdot\vec{E}\right)\right]. \tag{7.18}$$

The first term is a product of type 2 and requires the direct rule. The second term is of type 3, demanding the inverse rule, so that:

$$\begin{aligned}\left[\vec{D}\right] &= \left[\!\left[\varepsilon\right]\!\right]\left[\vec{E}_T\right] + \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\left[\vec{E}_N\right] \\ &= \left[\!\left[\varepsilon\right]\!\right]\left[\vec{E} - \vec{N}\left(\vec{N}.\vec{E}\right)\right] + \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\left[\vec{N}\left(\vec{N}.\vec{E}\right)\right].\end{aligned} \tag{7.19}$$

Introducing a square matrix representing a tensor product denoted $\left(\vec{N}\vec{N}\right)$ with elements given by N_iN_j, we obtain:

$$\left[\vec{D}\right] = \left[\!\left[\varepsilon\right]\!\right]\left[\vec{E}\right] + \left(\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} - \left[\!\left[\varepsilon\right]\!\right]\right)\left[\!\left[\vec{N}\vec{N}\right]\!\right]\left[\vec{E}\right] = Q_\varepsilon\left[\vec{E}\right], \tag{7.20}$$

where the matrix Q_ε has the form:

$$Q_\varepsilon = \left[\!\left[\varepsilon\right]\!\right] + \left(\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} - \left[\!\left[\varepsilon\right]\!\right]\right)\left[\!\left[\vec{N}\vec{N}\right]\!\right]. \tag{7.21}$$

In a similar manner for magnetic materials, we can find the link between magnetic field and induction in the truncated Fourier space:

$$\begin{aligned}&\left[\vec{B}\right] = Q_\mu \left[\vec{H}\right],\\ &\text{with}\quad Q_\mu = Q_\varepsilon = [\![\mu]\!] + \left(\left[\!\left[\frac{1}{\mu}\right]\!\right]^{-1} - [\![\mu]\!] \right) [\![\vec{N}\vec{N}]\!]\end{aligned} \tag{7.22}$$

Eq.(7.20) allows eliminating E_z in the system (7.6):

$$\begin{aligned}\left[E_z\right] &= Q_{\varepsilon,zz}^{-1}\left(\left[D_z\right] - Q_{\varepsilon,zx}\left[E_x\right] - Q_{\varepsilon,zy}\left[E_y\right]\right)\\ &= -Q_{\varepsilon,zz}^{-1}\left(\frac{\alpha\left[H_y\right] - \beta\left[H_x\right]}{\omega} + Q_{\varepsilon,zx}\left[E_x\right] + Q_{\varepsilon,zy}\left[E_y\right]\right)\end{aligned} \tag{7.23}$$

where the matrices α and β are defined in eq.(7.13).

Repeating the procedure for H_z:

$$\left[H_z\right] = Q_{\mu,zz}^{-1}\left(\frac{\alpha\left[E_y\right] - \beta\left[E_x\right]}{\omega} - Q_{\mu,zx}\left[H_x\right] - Q_{\mu,zy}\left[H_y\right]\right), \tag{7.24}$$

it is also eliminated from eqs. (7.6).

For **non-magnetic** media, the last expression is further simplified:

$$\left[H_z\right] = \frac{\alpha\left[E_y\right] - \beta\left[E_x\right]}{\omega\mu_0}. \tag{7.25}$$

Thus the system (7.6) is replaced by a system of ordinary differential equations:

$$\frac{d}{dz}\begin{pmatrix}\left[E_x\right]\\ \left[E_y\right]\\ \left[H_x\right]\\ \left[H_y\right]\end{pmatrix} = iM\begin{pmatrix}\left[E_x\right]\\ \left[E_y\right]\\ \left[H_x\right]\\ \left[H_y\right]\end{pmatrix}. \tag{7.26}$$

This equation can be expressed in a compressed form:

$$\frac{d}{dz}F(z) = iM(z)F(z) \tag{7.27}$$

Here the matrix M has 4x4 blocks:

$$
\begin{aligned}
M_{11} &= -\alpha Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zx} - Q_{\mu,yz} Q^{-1}_{\mu,zz} \beta \\
M_{12} &= -\alpha Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zy} + Q_{\mu,yz} Q^{-1}_{\mu,zz} \alpha \\
M_{13} &= -\omega Q_{\mu,yz} Q^{-1}_{\mu,zz} Q_{\mu,zx} + \frac{\alpha}{\omega} Q^{-1}_{\varepsilon,zz} \beta + \omega Q_{\mu,yx} \\
M_{14} &= -\omega Q_{\mu,yz} Q^{-1}_{\mu,zz} Q_{\mu,zy} - \frac{\alpha}{\omega} Q^{-1}_{\varepsilon,zz} \alpha + \omega Q_{\mu,yy} \\
M_{21} &= -\beta Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zx} + Q_{\mu,xz} Q^{-1}_{\mu,zz} \beta \\
M_{22} &= -\beta Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zy} - Q_{\mu,xz} Q^{-1}_{\mu,zz} \alpha \\
M_{23} &= \omega Q_{\mu,xz} Q^{-1}_{\mu,zz} Q_{\mu,zx} + \frac{\beta}{\omega} Q^{-1}_{\varepsilon,zz} \beta - \omega Q_{\mu,xx} \\
M_{24} &= \omega Q_{\mu,xz} Q^{-1}_{\mu,zz} Q_{\mu,zy} - \frac{\beta}{\omega} Q^{-1}_{\varepsilon,zz} \alpha - \omega Q_{\mu,xy} \\
M_{31} &= \omega Q_{\varepsilon,yz} Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zx} - \frac{\alpha}{\omega} Q^{-1}_{\mu,zz} \beta - \omega Q_{\varepsilon,yx} \\
M_{32} &= \omega Q_{\varepsilon,yz} Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zy} + \frac{\alpha}{\omega} Q^{-1}_{\mu,zz} \alpha - \omega Q_{\varepsilon,yy} \\
M_{33} &= -\alpha Q^{-1}_{\mu,zz} Q_{\mu,zx} - Q_{\varepsilon,yz} Q^{-1}_{\varepsilon,zz} \beta \\
M_{34} &= -\alpha Q^{-1}_{\mu,zz} Q_{\mu,zy} + Q_{\varepsilon,yz} Q^{-1}_{\varepsilon,zz} \alpha \\
M_{41} &= -\omega Q_{\varepsilon,xz} Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zx} - \frac{\beta}{\omega} Q^{-1}_{\mu,zz} \beta + \omega Q_{\varepsilon,xx} \\
M_{42} &= -\omega Q_{\varepsilon,xz} Q^{-1}_{\varepsilon,zz} Q_{\varepsilon,zy} + \frac{\beta}{\omega} Q^{-1}_{\mu,zz} \alpha + \omega Q_{\varepsilon,xy} \\
M_{43} &= -\beta Q^{-1}_{\mu,zz} Q_{\mu,zx} + Q_{\varepsilon,xz} Q^{-1}_{\varepsilon,zz} \beta \\
M_{44} &= -\beta Q^{-1}_{\mu,zz} Q_{\mu,zy} - Q_{\varepsilon,xz} Q^{-1}_{\varepsilon,zz} \alpha \,.
\end{aligned}
\tag{7.28}
$$

This form looks like the form of the M-matrix obtained by Lifeng Li for crossed anisotropic (electrically and magnetically) gratings with profiles invariant with respect to z [7.17].

Whatever the form of the matrix M, eq.(7.26) represents a linear set of first-order ordinary differential equations. It can be solved numerically (with several problems, discussed further on), using well developed numerical schemes. In the case of vertical invariance of the optogeometrical parameters of the system inside the modulated region, the elements of the M-matrix becomes constant in z, so that the solution of eq. (7.26) can be found through the eigenvectors and eigenvalues of M, a technique known under the name of Fourier modal method, or Rigorous coupled wave (RCW) method.

The solution of (7.26) gives a linear link between the field in the substrate and in the cladding

$$F(z_{max}) = T\, F(z_{min})\,, \tag{7.29}$$

where T is called transmission matrix.

The advantage of this presentation comes from the fact that the field components participating in the calculations are tangential to the interfaces between the substrate and the modulated region, and between the cladding and the modulated region, so that they are continuous across these interfaces (in the absence of surface charges).

7.3. Electromagnetic field in the homogeneous regions – plane wave expansion

In most case, the substrate and cladding are homogeneous isotropic media. The electromagnetic field there can be expressed as a sum of plane waves. In particular, if the x and y-dependencies are given as in eq.(7.7), the z-dependence is explicitly known, for example for the electric field it takes the form:

$$\vec{E}_p(z) = \vec{A}_p^{+} \exp(i\gamma_p z) + \vec{A}_p^{-} \exp(-i\gamma_p z) \tag{7.30}$$

of two waves propagating upwards (sign +) and downwards (sign –) along the z-axis, with p given in eq.(7.9). Each diffraction order with a given p propagates independently of the others, the coupling is effective inside the grating region.

The z-propagation constant γ depends on the medium properties:

$$\gamma_p = \sqrt{\omega^2 \varepsilon\mu - \alpha_p^2 - \beta_p^2}\,. \tag{7.31}$$

Equations (7.6) enable us to express the magnetic field components through the electric ones:

$$\begin{aligned} H_{x,p} &= -\frac{1}{\pm\gamma_p}\left(\frac{\alpha_p\beta_p}{\omega\mu}E_{x,p} + \frac{\beta_p^2+\gamma_p^2}{\omega\mu}E_{y,p}\right) \\ H_{y,p} &= \frac{1}{\pm\gamma_p}\left(\frac{\alpha_p^2+\gamma_p^2}{\omega\mu}E_{x,p} + \frac{\alpha_p\beta_p}{\omega\mu}E_{y,p}\right) \end{aligned} \tag{7.32}$$

where the sign of γ determines the direction of propagation in along z-axis.

With this link in mind, the column vector F in eq.(7.27) takes the form:

$$F \equiv \begin{pmatrix} [E_x] \\ [E_y] \\ [H_x] \\ [H_y] \end{pmatrix} = \Psi^{+}A^{+} + \Psi^{-}A^{-}, \tag{7.33}$$

where the column vectors

$$A^{\pm} = \begin{pmatrix} [A_x^{\pm}] \\ [A_y^{\pm}] \end{pmatrix} \tag{7.34}$$

contains the amplitudes of E_x and E_y propagating in positive or negative direction of the z-axis, matrices $\Psi^{\pm}$ are block-diagonal:

$$\Psi^{\pm} = \begin{pmatrix} \mathbb{I} & \\ & \mathbb{I} \\ \Psi^{\pm}_{xx} & \Psi^{\pm}_{xy} \\ \Psi^{\pm}_{yx} & \Psi^{\pm}_{yy} \end{pmatrix}, \tag{7.35}$$

with diagonal blocks

$$\begin{aligned} &\mathbb{I}_{pp} = 1, \\ &\Psi^{\pm}_{xx,pp} = \mp\frac{\alpha_p\beta_p}{\gamma_p\omega\mu}, \quad \Psi^{\pm}_{xy,pp} = \mp\frac{\beta_p^2+\gamma_p^2}{\gamma_p\omega\mu} \\ &\Psi^{\pm}_{yx,pp} = \pm\frac{\alpha_p^2+\gamma_p^2}{\gamma_p\omega\mu}, \quad \Psi^{\pm}_{yy,pp} = \pm\frac{\alpha_p\beta_p}{\gamma_p\omega\mu} \end{aligned} \tag{7.36}$$

found from eq.(7.32)

Let us consider the case of a single incident wave from the cladding. The grating generates different diffraction order that propagate upwards in the cladding and downwards in the substrate. We attribute number 1 to the substrate and number 3 to the cladding. The total number of unknown diffracted field amplitudes will be equal to $4P_{max}$, two sets of $A^{1-}_{x,p}$ and $A^{1-}_{y,p}$ transmitted in the substrate, and two sets of $A^{3+}_{x,p}$ and $A^{3+}_{y,p}$. These unknown amplitudes are subjected to $4P_{max}$ number of linear algebraic equation in (7.29).

In order to obtain the T-matrix, the numerical integration of eq.(7.26) is made by using the so-called shooting method, which consists of choosing $2P_{max}$ linearly independent representatives of the transmitted field. These representatives must correctly reflect the link between the electric and magnetic field components, as given by eqs.(7.32). A typical example for the shooting vectors starting from the substrate is that matrix Ψ^{1+}, which has $2P_{max}$ linearly independent columns. Here again, the number 1 indicates the substrate.

Thus the F column vector at $z = z_{min}$ can be formally written as a linear combination of the unknown amplitudes A^{1-}:

$$\tilde{F}(z_{min}) = \Psi^{1-}A^{1-}, \tag{7.37}$$

Assuming that there is no incidence from the substrate side. Here the tilde indicates that the vector F is not yet the true solution of the diffraction problem.

The result of the numerical integration from z_{min} to z_{max} will provide the values of $\tilde{F}$ at $z = z_{max}$, which are also a linear combination $\tilde{F}(z_{max})A^{1-}$ of A^{1-}, due to the linearity of the problem. On the other side, the column vector F at the upper interface is equal to $\psi^{3+}A^{3+} + \psi^{3-}A^{3-}$, according eq.(7.33), thus a linear set of algebraic equations for the unknown amplitudes A^{1-} and A^{3+} is obtained, with the free part determined by the wave incident from the cladding side:

$$\psi^{3+}A^{3+} + \psi^{3-}A^{3-} = \tilde{F}(z_{max})A^{1+}. \tag{7.38}$$

Once this system is solved, all field components can be calculated.

Unfortunately, this simple procedure creates enormous numerical problems that can be explained by using two different arguments:

First, it is known (but no quite well) in the theory of systems of ordinary differential equations, that numerical integration could become instable after a specific integration length, due to the fact that the set of shooting vectors can lose its linear independence during the integration. In other words, if the initial choice covers a vector space of $2P_{max}$ dimensions, this space could shrink during the numerical integration to reduce its dimensions, so that the final algebraic system (7.38) could become singular. A solution of the problem based on this understanding was proposed in 1990 by G. Tayeb by using intermediate orthonormalization procedures during the numerical integration.

The second argument is based on the fact that inside the modulated region, as well as in the homogeneous regions, electromagnetic field contains components that propagate both in the positive and in the negative z-direction. During the integration, they both are treated in the same manner. As far as the solution requires taking into account the evanescent orders in addition to the propagating ones, a part of the former grows exponentially in z-direction, while the other part decreases exponentially. Due to the limited length of computer words, the ones that decrease substantially will be lost with respect to the ones that grow rapidly, even if the former could bring physical information. During the 90s, several different algorithms were proposed for solving the problem, based on a different treatment of the diffraction orders propagating upwards and downwards [7.9, 7.10]. Among them, to so called S-matrix propagation algorithm [7.10c] is probably the easiest to implement. Moreover, it can be used with methods other than the differential one in, for example, treating a stack of layers by the integral method, or by methods based on a transformation of the coordinate system. Interested reader can find in Appendix 7.1 a brief description of the S-matrix algorithm.

7.4. Several simpler isotropic cases

In practice most applications use non-magnetic materials, for which the form of M-matrix is considerably simplified, taking into account that then Q_μ is diagonal and equal to μ_0. Furthermore, several specific cases are of great interest for application, and they lead to a further simplification of the M-matrix.

7.4.1. Classical grating with one-dimensional periodicity, example of sinusoidal profile

Let us consider a classical grating with grooves parallel to the y-axis and surface profile given by the equation z = g(x). The vector normal to the surface is given by

$$\begin{aligned} \vec{N} &= \frac{1}{\sqrt{1+g'^2(x)}}\left(-g'(x), 0, 1\right), \quad \text{if } g'(x) \text{ exists,} \\ \vec{N} &= (1,0,0), \quad \text{if not} \end{aligned} \qquad . \tag{7.39}$$

where the prime stands for a derivative with respect to x. In case of vertical walls $\vec{N} = (1, 0, 0)$. Thus the easiest way to generalize the normal vector to the entire modulated region is just to make it equal to eq.(7.39) not only on the profile z = g(x), but everywhere inside the grating region for $\min[g(x)] \le z \le \max[g(x)]$. The advantage of this choice is that $\vec{N}$ does not depend on z, and the Fourier transformation of the tensor $\vec{N}\vec{N}$ is done only once.

If the derivative of the profile function does not exist, or if the function is a multivalued one (e.g., circular or elliptical rods), but the interface can be expressed as a two-variable function:

$$g(x,y)=0\,, \tag{7.40}$$

the normal vector is easily defined as the gradient of the profile function $\vec{N}=\mathrm{grad}\left[g(x,z)\right]\Big/\left|\mathrm{grad}\left[g(x,z)\right]\right|$.

The Q_ε matrix takes the form:

$$Q_\varepsilon=\begin{pmatrix} [\![\varepsilon]\!]\left[\!\left[N_z^2\right]\!\right]+\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\left[\!\left[N_x^2\right]\!\right] & 0 & \left(\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}-[\![\varepsilon]\!]\right)[\![N_xN_z]\!] \\ 0 & [\![\varepsilon]\!] & 0 \\ \left(\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}-[\![\varepsilon]\!]\right)[\![N_xN_z]\!] & 0 & [\![\varepsilon]\!]\left[\!\left[N_x^2\right]\!\right]+\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\left[\!\left[N_z^2\right]\!\right] \end{pmatrix} \tag{7.41}$$

where it is taken into account that $N_x^2+N_z^2=1$. The fact that the normal vector components participate in the form of products in couples is important, because it leads to the conclusion is that the choice of the sign of $\vec{N}$ plays no role.

Further simplification of the M-matrix comes if limited to non-conical diffraction with $\beta_0 = 0$:

$$M=\begin{pmatrix} -\alpha Q_{\varepsilon,zz}^{-1}Q_{\varepsilon,zx} & 0 & 0 & -\frac{\alpha}{\omega}Q_{\varepsilon,zz}^{-1}\alpha+\omega\mu_0\mathbb{I} \\ 0 & 0 & -\omega\mu_0\mathbb{I} & 0 \\ 0 & \frac{\alpha\alpha}{\omega\mu_0}-\omega[\![\varepsilon]\!] & 0 & 0 \\ -\omega Q_{\varepsilon,xz}Q_{\varepsilon,zz}^{-1}Q_{\varepsilon,zx}+\omega Q_{\varepsilon,xx} & 0 & 0 & -Q_{\varepsilon,xz}Q_{\varepsilon,zz}^{-1}\alpha \end{pmatrix} \tag{7.42}$$

This shows that the system to integrate decouples into two subsystems, corresponding to the two fundamental polarizations, transversal with respect to the plane of incidence, transverse electric (TE):

$$\begin{aligned} \frac{d}{dz}\left[E_y\right]&=-i\omega\mu_0\left[H_x\right] \\ \frac{d}{dz}\left[H_x\right]&=i\left(\frac{\alpha^2}{\omega\mu_0}-\omega[\![\varepsilon]\!]\right)\left[E_y\right] \end{aligned} \tag{7.43}$$

and transverse magnetic (TM):

$$\begin{aligned} \frac{d}{dz}\left[E_x\right]&=-i\alpha Q_{\varepsilon,zz}^{-1}Q_{\varepsilon,zx}\left[E_x\right]-i\left(\frac{\alpha}{\omega}Q_{\varepsilon,zz}^{-1}\alpha-\omega\mu_0\right)\left[H_y\right] \\ \frac{d}{dz}\left[H_y\right]&=i\omega\left(Q_{\varepsilon,xx}-Q_{\varepsilon,xz}Q_{\varepsilon,zz}^{-1}Q_{\varepsilon,zx}\right)\left[E_x\right]-iQ_{\varepsilon,xz}Q_{\varepsilon,zz}^{-1}\alpha\left[H_y\right] \end{aligned} \tag{7.44}$$

7.4.1.1. Fourier transformation of the permittivity

The set of ordinary differential equations to be integrated contains the Fourier transforms of ε, 1/ε, μ, 1/μ, N_x^2, and N_z^2. In general, Fast Fourier transform (FFT) techniques can be easily applied. As already discussed with respect to eq.(7.39), the normal vector components must be transformed only once, if chosen to be independent on z. On the other hand, the permittivity and permeability depend on z and their Fourier components have to be calculated for each value of z during the numerical integration. Fortunately, in the 1D case, it is possible and recommended to use analytical formulae for the Fourier transforms of ε, 1/ε, μ, 1/μ, which give faster more accurate results. This can be done because for a given value of z, they are piecewise constant functions of y. Fig.7.3 presents schematically a grating with a period d that separates two homogeneous media with permittivities ε_1 and ε_3. For a given value z_0 of z, the Fourier transform of, for example, the permittivity inside the modulated region $0 \le z \le h$ is given by

$$\begin{aligned}\varepsilon_m &= \frac{\varepsilon_1}{d}\int_{x_1}^{x_2} e^{-imK_x x}dx + \frac{\varepsilon_3}{d}\int_{x_2}^{d+x_1} e^{-imK_x x}dx \\ &= (\varepsilon_1-\varepsilon_3)\frac{\sin\left(mK_x\dfrac{x_2-x_1}{2}\right)}{\pi m}e^{-imk_x\frac{x_1+x_2}{2}} + \varepsilon_3\delta_{m,0}\end{aligned} \tag{7.45}$$

so that the two integrals can be solved analytically, once x_1 and x_2 are determined from the inverse of g(x):

$$x_{1,2} = g^{-1}(z_0)\,. \tag{7.46}$$

If the inverse of g(x) has more than two solutions, the sum of integrals (7.45) will contain several more terms. The same equations can be used to obtain the Fourier transforms of the inverse of the permittivity.

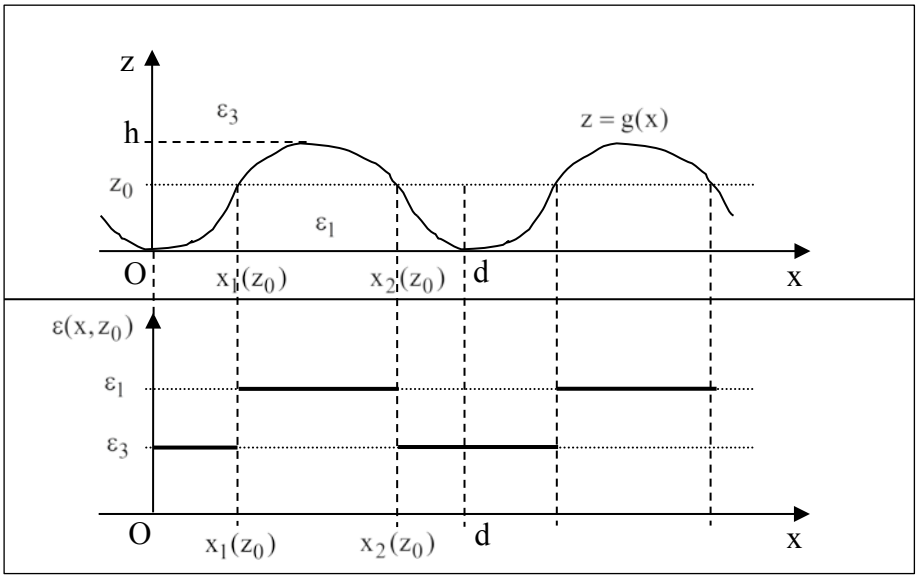


Fig.7.3.Piecewise constant representation of the permittivity for a one-dimensional grating

In the case of a sinusoidal profile:

$$z = \frac{h}{2}\left[1 + \sin(K_x x)\right] \quad \Rightarrow \quad x_{1,2} = \arcsin\left(\frac{2z_0}{h} - 1\right). \tag{7.47}$$

7.4.1.2. Fourier transformation of the normal vector

As already explained, the Fourier transformation of the normal vector requires its continuation all over the space. If the grating profile can be represented as a single-value function, we can use eq.(7.39) for $\vec{N}$ and calculate the Fourier components of the tensor $\vec{N}\vec{N}^T$ by use of the Fast Fourier transform (FFT) technique once for all z-values. For a sinusoidal grating having a profile defined in eq.(7.47), the normal vector takes the form:

$$\vec{N} = \frac{1}{\sqrt{1 + g'^2(x)}}\left(-g'(x), 0, 1\right) = \frac{\left(-\frac{\pi h}{d}\cos(K_x x), 0, 1\right)}{\sqrt{1 + \left(\frac{\pi h}{d}\right)^2 \cos^2(K_x x)}} \tag{7.48}$$

.

7.4.2. Classical isotropic trapezoidal or triangular grating

A trapezoidal grating is shown schematically in Fig.7.4 with two flat regions L at the top and the bottom of the groove and two different, in general, groove angles ψ. The Fourier transform of the permittivity and its inverse are calculated using eq.(7.45) with:

$$\begin{aligned} x_1 &= z_0 \mathrm{cotg}\psi_1 \\ x_2 &= x_C - z_0 \mathrm{cotg}\psi_2 \end{aligned} \tag{7.49}$$

with $x_C = d - L_2$. For the normal vector, the period can be divided in four regions A to D, as shown in the figure:

$$\begin{aligned} &N_y = 0 \\ &\left.\begin{aligned} N_x &= \sin\psi_1 \\ N_z &= -\cos\psi_1 \end{aligned}\right\} \text{ in A,} \qquad \left.\begin{aligned} N_x &= \sin\psi_2 \\ N_z &= \cos\psi_2 \end{aligned}\right\} \text{ in C,} \\ &\left.\begin{aligned} N_x &= 1 \\ N_z &= 0 \end{aligned}\right\} \text{ in B and D,} \end{aligned} \tag{7.50}$$

Their Fourier components do not depend on z and can be represented as a sum of several analytical terms, similar to eq.(7.45):

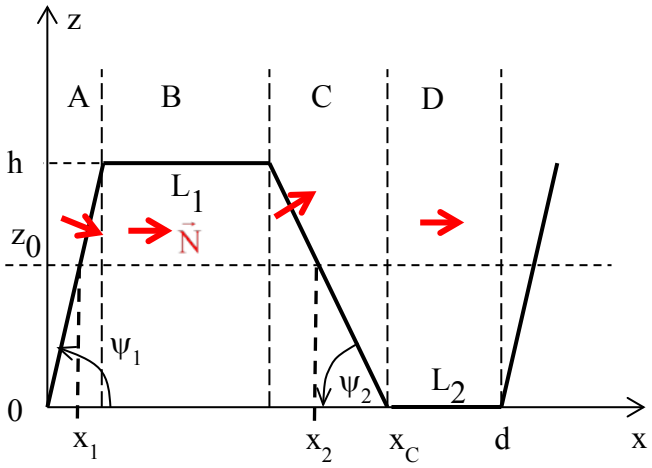


Fig.7.4. Trapezoidal profile with parameters. The normal vector direction is given in red arrows.

$$\left(N_x^2\right)_m = \frac{\sin^2\psi_1}{d}\int_0^{h\,\mathrm{cotg}\psi_1} e^{-imK_x x}dx + \frac{1}{d}\int_{h\,\mathrm{cotg}\psi_1}^{h\,\mathrm{cotg}\psi_1+L_1} e^{-imK_x x}dx + \frac{\sin^2\psi_2}{d}\int_{h\,\mathrm{cotg}\psi_1+L_1}^{d-L_2} e^{-imK_x x}dx + \frac{1}{d}\int_{d-L_2}^{d} e^{-imK_x x}dx \tag{7.51}$$

$$\left(N_y^2\right)_m = \frac{\cos^2\psi_1}{d}\int_0^{h\,\mathrm{cotg}\psi_1} e^{-imK_x x}dx + \frac{\cos^2\psi_2}{d}\int_{h\,\mathrm{cotg}\psi_1+L_1}^{d-L_2} e^{-imK_x x}dx \tag{7.52}$$

$$\left(N_x N_y\right)_m = -\frac{\sin 2\psi_1}{2d}\int_0^{h\,\mathrm{cotg}\psi_1} e^{-imK_x x}dx + \frac{\sin 2\psi_2}{2d}\int_{h\,\mathrm{cotg}\psi_1+L_1}^{d-L_2} e^{-imK_x x}dx. \tag{7.53}$$

A triangular-groove grating can be considered as a particular case of a trapezoidal profile with no flat regions, $L_1 = L_2 = 0$, $x_C = d$. Moreover, the profile given in Fig.7.4 also includes the case with vertical facets, and some more exotic profiles with hanging back walls, Fig.7.5.

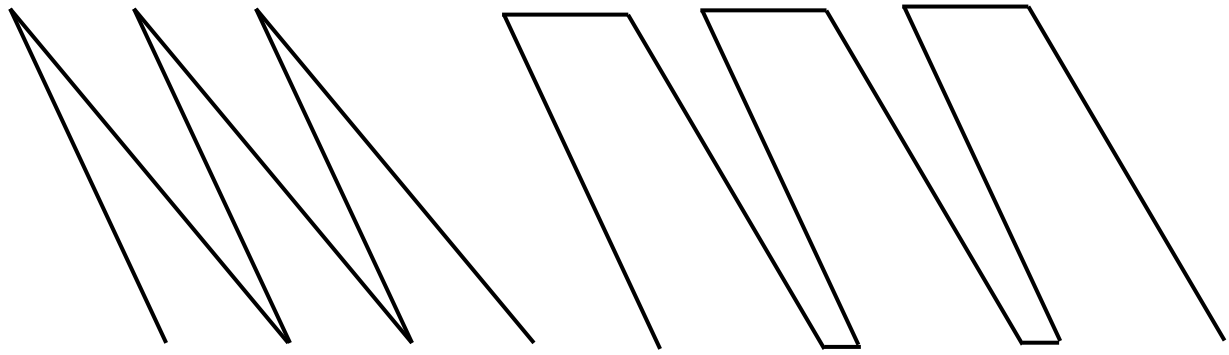


Fig.7.5. Two different profiles with slanted grooves

7.4.3. Classical lamellar grating

Lamellar profile with vertical walls is most easy to treat, because the normal to the profile vector has only one non-zero component, $N_x = 1$. The Q_ε matrix takes the form:

$$Q_\varepsilon = \begin{pmatrix} \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} & 0 & 0 \\ 0 & [\![\varepsilon]\!] & 0 \\ 0 & 0 & [\![\varepsilon]\!] \end{pmatrix} \tag{7.54}$$

$$M = \begin{pmatrix} 0 & 0 & \frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\beta_0 & -\frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\alpha + \omega\mu_0 \mathbb{I} \\ 0 & 0 & \frac{\beta_0^2}{\omega}[\![\varepsilon]\!]^{-1} - \omega\mu_0\mathbb{I} & -\frac{\beta_0}{\omega}[\![\varepsilon]\!]^{-1}\alpha \\ -\frac{\alpha}{\omega\mu_0}\beta_0 & \frac{\alpha^2}{\omega\mu_0} - \omega[\![\varepsilon]\!] & 0 & 0 \\ \omega\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} - \frac{\beta_0^2}{\omega\mu_0}\mathbb{I} & \frac{\beta_0}{\omega\mu_0}\alpha & 0 & 0 \end{pmatrix} \tag{7.55}$$

In non-conical diffraction, when $\beta_0 = 0$, the two fundamental polarizations are decoupled and can be solved independently of each other. The M-matrix is simplified to obtain an antidiagonal block form:

$$M = \begin{pmatrix} 0 & 0 & 0 & -\frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\alpha + \omega\mu_0\mathbb{I} \\ 0 & 0 & -\omega\mu_0\mathbb{I} & 0 \\ 0 & \frac{\alpha^2}{\omega\mu_0} - \omega[\![\varepsilon]\!] & 0 & 0 \\ \omega\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} & 0 & 0 & 0 \end{pmatrix} \tag{7.56}$$

thus the two sets of differential equations for each polarization become:

$$\begin{aligned} \frac{d}{dz}[E_x] &= i\left(\omega\mu_0\mathbb{I} - \frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\alpha\right)[H_y] \\ \frac{d}{dz}[H_y] &= i\omega\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}[E_x] \end{aligned} \tag{7.57}$$

and

$$\begin{aligned} \frac{d}{dz}[E_y] &= -i\omega\mu_0[H_x] \\ \frac{d}{dz}[H_x] &= i\left(\frac{\alpha^2}{\omega\mu_0} - \omega[\![\varepsilon]\!]\right)[E_y] \end{aligned} \tag{7.58}$$

Even in the case of conical diffraction, it is possible to define two other polarizations, for which the differential system decouples. These are the electric and magnetic polarizations that are *transverse with respect to the x-axis*. Let us denote the two polarization with superscripts (e), when $E_x = 0$, and (h), when $H_x = 0$. For (e) case, it is possible to express H_y as a function of H_x from eq.(7.26) and the first line of the M-matrix in eq.(7.55):

$$\left[H_y^{(e)}\right] = -\left(\omega\mu_0 \mathbb{I} - \frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\alpha\right)^{-1}\frac{\alpha}{\omega}[\![\varepsilon]\!]^{-1}\beta_0\left[H_x^{(e)}\right] \tag{7.59}$$

which can be simplified into:

$$\left[H_y^{(e)}\right] = -\alpha\left(\omega^2\mu_0[\![\varepsilon]\!] - \alpha^2\right)^{-1}\beta_0\left[H_x^{(e)}\right] \tag{7.60}$$

The second line of the matrix M then results in:

$$\frac{d}{dz}\left[E_y^{(e)}\right] = i\left[\frac{\beta_0^2}{\omega}[\![\varepsilon]\!]^{-1} - \omega\mu_0\mathbb{I} + \frac{\beta_0^2}{\omega}[\![\varepsilon]\!]^{-1}\alpha^2\left(\omega^2\mu_0[\![\varepsilon]\!] - \alpha^2\right)^{-1}\right]\left[H_x^{(e)}\right] \tag{7.61}$$

This expression can be further simplified, and together with the third line of eq.(7.55) (when $E_x = 0$) gives a set of equation for (e) polarization:

$$\begin{aligned}\frac{d}{dz}\left[E_y^{(e)}\right] &= i\omega\mu_0\left[\beta_0^2\left(\omega^2\mu_0[\![\varepsilon]\!] - \alpha^2\right)^{-1} - \mathbb{I}\right]\left[H_x^{(e)}\right]\\ \frac{d}{dz}\left[H_x^{(e)}\right] &= \frac{i}{\omega\mu_0}\left(\alpha^2 - \omega^2\mu_0[\![\varepsilon]\!]\right)\left[E_y^{(e)}\right]\end{aligned} \tag{7.62}$$

Similar procedure for (h) case when $H_x^{(h)} = 0$, result in another system of differential equations, decoupled from the (e) case:

$$\begin{aligned}\frac{d}{dz}\left[H_y^{(h)}\right] &= i\omega\left[\left[\!\!\left[\frac{1}{\varepsilon}\right]\!\!\right]^{-1} + \beta_0^2\left(\alpha[\![\varepsilon]\!]^{-1}\alpha - \omega^2\mu_0\mathbb{I}\right)^{-1}\right]\left[E_x^{(h)}\right]\\ \frac{d}{dz}\left[E_x^{(h)}\right] &= \frac{i}{\omega}\left(\omega^2\mu_0\mathbb{I} - \alpha[\![\varepsilon]\!]^{-1}\alpha\right)\left[H_y^{(h)}\right]\end{aligned} \tag{7.63}$$

In non-conical mount, $\beta_0 = 0$ and eqs.(7.62) and (7.63) become equivalent to eqs.(7.58) and (7.57).

Both conical and nonconical cases of diffraction by lamellar gratings are solved by eigenvector technique, due to the fact that the coefficients of the differential equations are z-independent. Moreover, due to the separation of the two fundamental polarizations, it is possible to further reduce by half the size of the matrices, by dealing with second-order differential equations. For example, eq. (7.57) can be written in the form:

$$\begin{aligned}\frac{d}{dz}\left[E_x\right] &= iM_{14}\left[H_y\right]\\ \frac{d}{dz}\left[H_y\right] &= iM_{41}\left[E_x\right]\end{aligned} \tag{7.64}$$

Thus

$$\begin{aligned}\frac{d^2}{dz^2}\left[E_x\right] &= -M_{14}M_{41}\left[E_x\right]\\ \frac{d^2}{dz^2}\left[H_y\right] &= -M_{41}M_{14}\left[H_y\right]\end{aligned} \tag{7.65}$$

Les us denote with ρ_p^2 the eigenvalues of the product $M_{14}M_{41}$ and with V the matrix with its eigenvectors arranged in columns. The solution of the first eq.(7.65) can be written as:

$$\left[E_x(z)\right] = V\Phi(z)V^{-1}\left[E_x(0)\right] \tag{7.66}$$

with

$$\Phi_{pp'}(z) = \delta_{pp'}\exp(\pm i\rho_p z) \tag{7.67}$$

which shows that the elementary solutions along z (called *modes*, wherefrom the names *Fourier modal method* or *Rigorous coupled waves method*) exist in pairs that can propagate upwards or downwards with the same propagation constants.

By integrating the second eq.(7.65), we obtain that:

$$\left[H_y(z)\right] = W\Phi(z)W^{-1}\left[H_y(0)\right] \tag{7.68}$$

with W that can be written in different forms, because the eigenvectors are defined within an arbitrary factor. For example, if we take into account the second eq.(7.64), $W = \mp i\rho^{-1}M_{41}V$, where the diagonal matrix ρ has elements equal to ρ_p. Another possibility, that is quite convenient in TM polarization described by eq.(7.57), is at first to calculate the eigenvectors of $M_{41}M_{14}$, instead of $M_{14}M_{41}$ (their eigenvalues are the same). Then the link between V and W contains the inverse of M_{41}, which is just equal to $\frac{1}{\omega}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]$, so that $W = \pm\frac{i}{\omega}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]V\rho$.

7.4.4. Crossed grating having vertical walls made of isotropic material

Most of the recent applications of the Fourier modal method are devoted to studies of light diffraction by structures with 2D periodicity and piecewise invariant in the third direction. This popularity has several reasons. First, extraordinary light transmission was found in the late 90s by Ebbesen [7.18] for such structures, namely metallic sheets with periodical hole arrays, and it attracted a lot of attention (see Chapter 1). Second, the technology of such structures has significantly advanced in the last 20 years. Third, the Fourier modal method is relatively simple to implement, and much faster than most of the other methods, because of the eigenvalue/vector technique of integration.

Detailed study of these structures is described in Chapter 13. However, due to its importance, we are discussing different aspects of this theory, as it presents a particular case of the more general geometry, that is characterized by a constant value of the z-component of the normal vector on each cross-section having z = const. The prolongation of the normal vector within the entire grating cell is discussed in sections 7.6 and 7.7, in which a detailed description of analytical Fourier transform in given for inclusions with elliptical cross-section.

7.5. Differential theory for anisotropic media

If we consider anisotropic media that do not extend inside the grating structure, there is not necessary to reformulate the diffraction theory, only that in the general case it is not possible to separate the problem into two independent polarizations, and it is necessary to work with the complete 4Pmax vectors and matrices.

In the case of anisotropic medium that lies inside the grating, the equations linking the M-matrix with the Q_ε and Q_μ matrices remain the same, eq.(7.28) for isotropic and anisotropic media. The difference comes from the fact that the Q-matrices take more complex form, because the link between the normal and tangential components of the couples E and D and H and B is made through the tensors of permittivity and permeability, which are not scalars. Let us establish this link in detail for E and D. As far as the continuous and discontinuous field components must be factorized differently, we construct a column vector F_ε, which contains the continuous field components E_T and D_N. There are two tangential components to the grating surface, and only a single normal:

$$F_\varepsilon = \begin{pmatrix} D_N \\ E_{T_1} \\ E_{T_2} \end{pmatrix} = \begin{pmatrix} \vec{N}\cdot\left(\overline{\overline{\varepsilon}}\,\vec{E}\right) \\ \vec{T}_1\cdot\vec{E} \\ \vec{T}_2\cdot\vec{E} \end{pmatrix} = U_\varepsilon \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \tag{7.69}$$

where the double bar indicates a second-rank tensor with 3 dimensions, and the matrix U_ε has the form:

$$U_\varepsilon = \begin{pmatrix} \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)_x & \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)_y & \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)_z \\ T_{1x} & T_{1y} & T_{1z} \\ T_{2x} & T_{2y} & T_{2z} \end{pmatrix} \tag{7.70}$$

with $\vec{N}\,\overline{\overline{\varepsilon}}$ representing a tensor product with contraction of indices, for example, $\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)_x = N_x\overline{\overline{\varepsilon}}_{xx} + N_y\overline{\overline{\varepsilon}}_{yx} + N_z\overline{\overline{\varepsilon}}_{zx}$, etc.

The vectors $\vec{N}, \vec{T}_1, \text{and}\, \vec{T}_2$ are defined on the grating surface, but for their further Fourier transform, it is necessary to choose a suitable continuation. The necessary conditions are that (i) they are continuous on the surfaces where ε and μ are discontinuous, and (ii) they form an orthonormal triad.

Since $\overline{\overline{\varepsilon}}$ never vanishes, the determinant of U_ε represents a quadric non-null form, equal to:

$$\xi_\varepsilon \equiv \det U_\varepsilon = \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\cdot\left(\vec{T}_1\times\vec{T}_2\right) = \sum_{i,j=x,y,z} N_i\varepsilon_{ij}N_j \tag{7.71}$$

since $\vec{N} = \vec{T}_1\times\vec{T}_2$.

Thus U_ε has an inverse U_ε^{inv} in the form:

$$U_\varepsilon^{inv} = \frac{1}{\xi_\varepsilon}\begin{pmatrix} N_x & -\left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2\right]_x & \left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_1\right]_x \\ N_y & -\left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2\right]_y & \left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_1\right]_y \\ N_z & -\left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2\right]_z & \left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_1\right]_z \end{pmatrix} \tag{7.72}$$

It is not evident to derive this form, but it can easily be verified by using the equivalence $U_\varepsilon U_\varepsilon^{inv} = \mathbb{I}$ and the fact that $\vec{N} = \vec{T}_1 \times \vec{T}_2$. For example, the product of the second line of U_ε with the second column of U_ε^{inv} can be written in vectorial form:

$$\xi_\varepsilon\left(U_\varepsilon U_\varepsilon^{inv}\right)_{yy} = -\vec{T}_1\cdot\left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2\right] = -\left[\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2\right]\cdot\vec{T}_1 = -\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\cdot\left(\vec{T}_2\times\vec{T}_1\right) = \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\cdot\vec{N} = \xi_\varepsilon \tag{7.73}$$

Going back to the vector F_ε, it is continuous across the grating surface, whereas the Cartesian components of the electric vector are, in general discontinuous, as well as the components of U_ε. Thus for their Fourier transform, it is necessary to apply the inverse factorization rule:

$$\left[F_\varepsilon\right] = \left[\!\left[U_\varepsilon^{inv}\right]\!\right]^{-1}\left[\vec{E}\right] \tag{7.74}$$

At the other hand,

$$\vec{E} = U_\varepsilon^{inv}F_\varepsilon \quad\Rightarrow\quad \vec{D} = \overline{\overline{\varepsilon}}\,U_\varepsilon^{inv}F_\varepsilon \tag{7.75}$$

with F_ε being continuous, so that the Fourier transform of $\vec{D}$ requires the direct factorization rule:

$$\left[\vec{D}\right] = \left[\!\left[\overline{\overline{\varepsilon}}\,U_\varepsilon^{inv}\right]\!\right]\left[\!\left[U_\varepsilon^{inv}\right]\!\right]^{-1}\left[\vec{E}\right] \tag{7.76}$$

i.e.,

$$Q_\varepsilon = \left[\!\left[\overline{\overline{\varepsilon}}\,U_\varepsilon^{inv}\right]\!\right]\left[\!\left[U_\varepsilon^{inv}\right]\!\right]^{-1} \tag{7.77}$$

For gratings having anisotropic magnetic properties, the corresponding Q_μ matrix is obtained from eqs. (7.71), (7.72), and (7.77) by replacing U_ε^{inv} by U_μ^{inv} and $\overline{\overline{\varepsilon}}$ by $\overline{\overline{\mu}}$.

7.5.1. Lamellar gratings made of anisotropic material

Such gratings are analyzed in Chapter 13, devoted to the Fourier modal method by using more direct approaches, but here we want show how the corresponding equations can be obtained from the general eqs.(7.72). To this aim it is sufficient to realize that

$$\begin{aligned} \vec{N} &= (1,0,0) \\ \vec{T}_1 &= (0,1,0) \\ \vec{T}_2 &= (0,0,1) \end{aligned} \tag{7.78}$$

so that

$$\begin{aligned}\left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_2 &= (\varepsilon_{xy}, -\varepsilon_{xx}, 0)\\ \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\times\vec{T}_1 &= (-\varepsilon_{xz}, 0, \varepsilon_{xx})\\ \left(\vec{N}\,\overline{\overline{\varepsilon}}\right)\cdot\vec{N} &= \varepsilon_{xx}\end{aligned} \tag{7.79}$$

and eq.(7.72) takes the form:

$$U_\varepsilon^{inv} = \begin{pmatrix} \dfrac{1}{\varepsilon_{xx}} & -\dfrac{\varepsilon_{xy}}{\varepsilon_{xx}} & -\dfrac{\varepsilon_{xz}}{\varepsilon_{xx}} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7.80}$$

with a determinant equal to $\left[\!\left[\dfrac{1}{\varepsilon_{xx}}\right]\!\right]$. Thus

$$\left[\!\left[U_\varepsilon^{inv}\right]\!\right]^{-1} = \begin{pmatrix} \left[\!\left[\dfrac{1}{\varepsilon_{xx}}\right]\!\right]^{-1} & \left[\!\left[\dfrac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\dfrac{\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right] & \left[\!\left[\dfrac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\dfrac{\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right] \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathbb{I} \end{pmatrix} \tag{7.81}$$

The second matrix that is required takes the form:

$$\overline{\overline{\varepsilon}}U_\varepsilon^{inv} = \begin{pmatrix} 1 & 0 & 0 \\ \dfrac{\varepsilon_{yx}}{\varepsilon_{xx}} & \varepsilon_{yy} - \dfrac{\varepsilon_{yx}\varepsilon_{xy}}{\varepsilon_{xx}} & \varepsilon_{yz} - \dfrac{\varepsilon_{yx}\varepsilon_{xz}}{\varepsilon_{xx}} \\ \dfrac{\varepsilon_{zx}}{\varepsilon_{xx}} & \varepsilon_{zy} - \dfrac{\varepsilon_{zx}\varepsilon_{xy}}{\varepsilon_{xx}} & \varepsilon_{zz} - \dfrac{\varepsilon_{zx}\varepsilon_{xz}}{\varepsilon_{xx}} \end{pmatrix} \tag{7.82}$$

This form is valid even when the permittivity tensor is not symmetric, as happens in the modeling of magnetooptical effects.

The Q_ε matrix takes the form obtained in [7.19], using a completely different approach:

$$Q_\varepsilon = \left[\!\left[\overline{\overline{\varepsilon}}\,U_\varepsilon^{inv}\right]\!\right]\left[\!\left[U_\varepsilon^{inv}\right]\!\right]^{-1} \tag{7.83}$$

$$= \begin{pmatrix} \left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1} & \left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right] & \left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right] \\ \left[\!\left[\frac{\varepsilon_{yx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1} & \left[\!\left[\varepsilon_{yy}-\frac{\varepsilon_{yx}\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right]+\left[\!\left[\frac{\varepsilon_{yx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right] & \left[\!\left[\varepsilon_{yz}-\frac{\varepsilon_{yx}\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right]+\left[\!\left[\frac{\varepsilon_{yx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right] \\ \left[\!\left[\frac{\varepsilon_{zx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1} & \left[\!\left[\varepsilon_{zy}-\frac{\varepsilon_{zx}\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right]+\left[\!\left[\frac{\varepsilon_{zx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xy}}{\varepsilon_{xx}}\right]\!\right] & \left[\!\left[\varepsilon_{zz}-\frac{\varepsilon_{zx}\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right]+\left[\!\left[\frac{\varepsilon_{zx}}{\varepsilon_{xx}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon_{xx}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon_{xz}}{\varepsilon_{xx}}\right]\!\right] \end{pmatrix}$$

7.6. Normal vector prolongation for 2D periodicity; Fourier transform

As observed, the proper use of the direct and the inverse factorization rules requires that the vector normal to the interfaces between different media is defined not only on these interfaces, but throughout the entire grating cell. In the case of classical gratings with one-dimensional periodicity, the prolongation of the normal vector can be done quite easily, as shown in sec.7.4.1. For two-dimensional periodicity, the choice depends on the geometry, but also on its mathematical representation. Several different solutions have to be considered, without pretending to be exhaustive.

In general, the cross-section profile changes with z, so that the matrices Q_ε, Q_μ, and M have to be recalculated for each value of z. If the geometry is z-invariant, this must be done only once. Concerning the Fourier components of the normal vector, there are two different classes of grating profiles that has to be treated separately. The first class consists of surfaces that can be expressed all over the unit cell (containing a single period in x and z) as an analytical (at lease piecewisely) function $z_S = g(x, y)$, where S indicates that the point lies on the interface. In this case, it is possible to have a unique extension of $\vec{N}$ whatever the values of z. In addition, it is not necessary to calculate the cross-section of the surface(s) with a plane perpendicular to the z-axis for each value of z. This case also includes multilayered homomorphous structure with constant layer(s) thickness in the z-direction. We consider this class of cases in sec.7.6.1.

The second class of surfaces includes surfaces that cannot be expressed through single-valued functions, as the example given on the right-hand side of Fig.7.12. In that case, it is necessary for each fixed value of z to know the cross-section function of the grating surface with the plane z = const. Subsection 7.6.2 presents general analysis, some important specific cases are considered further in the following subsections.

7.6.1. General analytical surfaces

If the interface representing the structure can be expressed as a single-valued function, analytical over the entire unit cell (this is also valid if different analytical functions can be defined over different regions of the cell):

$$z_S = g(x, y), \tag{7.84}$$

then the components of the vector normal to the surface defined on the surface have the form:

$$\vec{N}(x,y)=\frac{\left(-\frac{\partial g}{\partial x},-\frac{\partial g}{\partial y},1\right)}{\sqrt{1+\mathrm{grad}^2 g(x,y)}}. \tag{7.85}$$

It can immediately be extended to whatever the values of z inside the modulated region. Moreover, its values do not depend on z, as it was a case of the classical one-dimensional grating already discussed in sec.7.4.1.

The permittivity and its inverse can easily be obtained on a mesh (x,y) covering the grating cell for each z:

$$\begin{aligned} g(x,y)<z \quad &\Rightarrow \quad \varepsilon(x,y)=\varepsilon_3 \\ g(x,y)\geq z \quad &\Rightarrow \quad \varepsilon(x,y)=\varepsilon_1 \end{aligned} \tag{7.86}$$

where 1 and 3 are the indexes representing respectively the inferior and the superior regions separated by the grating surface (as it was done in Fig.7.3). If the cross-section of the grating surface with the planes z = const are ellipses (or circles), and if N_z does not depend on (x,y) at each z, it is possible to replace the numerical Fourier transform by an analytical formulae, as discussed in Sec.7.7. One important particular case is the z-invariant grating with elliptical cross section; another case includes the gratings having a rotational symmetry, as shown in Fig.7.12.

The same extension (7.85) for the normal vector is valid for a stack of layers having homogeneous thicknesses in the z-direction:

$$z_{S,j}=g(x,y)+t_j \tag{7.87}$$

The permittivity and its inverse inside the intermediate layers are simply given as:

$$z_{S,j-1}(x,y)<z\leq z_{S,j}(x,y) \quad \Rightarrow \quad \varepsilon(x,y)=\varepsilon_j. \tag{7.88}$$

7.6.2. Irregular general surfaces

If the case does not fit into the preceding section, the interface is expressed through the more general function $u(x,y,z_S)=0$, and the vector $\vec{N}$ has to be determined for each inclusion:

$$\vec{N}(x,y,z_S)=\frac{\left(\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\frac{\partial u}{\partial z_S}\right)}{\left|\mathrm{grad}\ u(x,y,z_S)\right|} \tag{7.89}$$

However, these values are well defined on the grating surface (except on its edges), and have to be extended over the entire cell. When considering a cross-section of the profile with a plane at z = const., several different cases exist:

7.6.2.1. Single-valued radial cross-section

At first, we shall consider that the cross section function $f(x_S,y_S)=0$ defines a single curve, and that curve can be expressed in cylindrical coordinates as

$$\rho_S = \rho_S(\varphi) \tag{7.90}$$

where

$$\begin{aligned} \rho_S &= \sqrt{(x_S - x_C)^2 + (y_S - y_C)^2} \\ \varphi &= \arctan\left[(y_S - y_C)/(x_S - x_C)\right] \end{aligned} \tag{7.91}$$

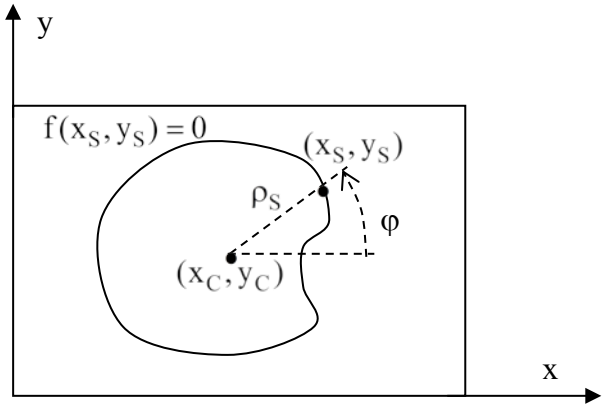


Fig.7.6. Single-curve cross-section of the grating surface at z = const.

and x_C and y_C represent a "central" point of the curve, Fig.7.6. Here we assume that the values of ρ_S are unique for each φ. The other case is analyzed further on.

It is possible to extend to the entire cell the values of the normal vector, defined only on the curve, by assuming that it is constant for each fixed angle φ. This prolongation requires the following procedure:

1. Fixing the pair (x,y).
2. Calculating the angle $\varphi = \arctan\left[(y - y_C)/(x - x_C)\right]$.
3. Calculating $\rho_S = \rho_S(\varphi)$ from eq.(7.91).
4. Calculation of $x_S = \rho_S \cos\varphi$, and $y_S = \rho_S \sin\varphi$.
5. Determining N_z, together with N_x and N_y from eq.(7.89).
6. Attributing these values of the components of $\vec{N}(x_S, y_S)$ to the pair (x,y).
7. Fast Fourier transform after the normal vector components are determined for all the pairs (x,y) on a mesh inside the grating cell.

The procedure can be simplified for most of the typical diffracting objects, as shown further for objects with elliptical or circular cross-section.

If the grating profile varies with z, the calculations of the Fourier components of the permittivity and its inverse has to be made at each value of z, both for the analytical profiles, for which the normal vector prolongation can be chosen z-invariant, or for the irregular surfaces. For each (x,y) pair of the mesh used in the FFT method, it is possible to determine whether the point lies within or outside the cross-section part of Fig.7.6:

$$\begin{aligned} &\rho(\varphi) < \rho_S(\varphi), \quad \varepsilon(x,y) = \varepsilon_{\text{inside}} \\ &\rho(\varphi) \geq \rho_S(\varphi), \quad \varepsilon(x,y) = \varepsilon_{\text{outside}} \end{aligned} \tag{7.92}$$

with $\rho = \sqrt{(x - x_C)^2 + (y - y_C)^2}$.

7.6.2.2. Objects with polygonal cross section

A typical example of such objects is presented in Fig.7.1. Its surface consists of different plates, and for their treatment the condition N_z = const. for fixed z is fulfilled, because $\vec{N}$ is constant at each plate. Another possible surface consists of plane ribbons with curvature in z-direction, Fig.7.7.

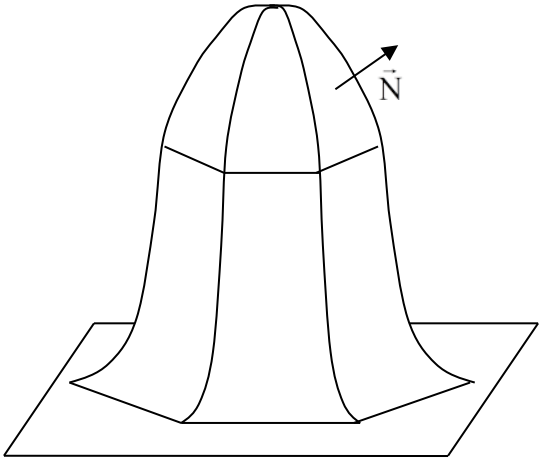


Fig.7.7. Surface made of plane ribbons

As shown in Fig.7.8, the cross-section represents a polygon. On each of its sides the modulus of the in-plane component of the normal vector is known:

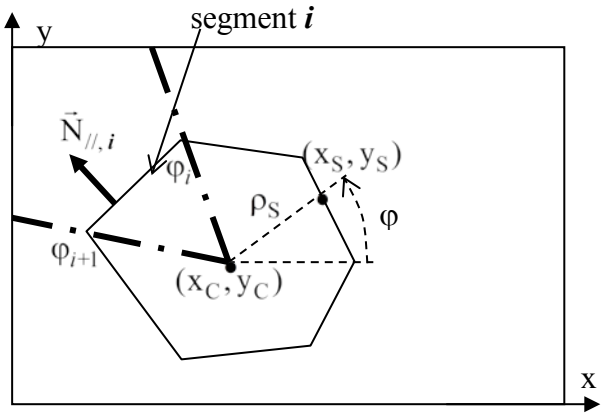


Fig.7.8. Object with a polygonal cross-section.

$$N_{//,i} = \sqrt{1 - N_{z,i}^2} \tag{7.93}$$

and its direction is perpendicular to the segment. If the *i*-th segment is located between the angles φ_i and φ_{i+1}, we can extend the definition of the normal vector all over the unit cell situated within the range (φ_i, φ_{i+1}), delimited by the bold dot-dashed lines in Fig.7.8, by

assuming that $\vec{N} = \vec{N}_i$. The normal vector extension will be continuous everywhere, except on the sector border lines (bold dot-dashed lines) and thus the only points where both permittivity (and/or permeability) and $\vec{N}$ are simultaneously discontinuous are at the polygonal corners, where anyway $\vec{N}$ is never continuous.

The procedure to follow requires that for each value of z the polygon corners (x_i, y_i) and the z-components of $\vec{N}$ for each segment are determined, as well as fixing some "central" point (x_C, y_C). Then the angular ranges of each segment with respect to that central point are calculated:

$$\varphi_i = \arctan\frac{y_i - y_C}{x_i - x_C} \tag{7.94}$$

For each pair (x,y), the azimuthal angle is given as $\varphi = \arctan\left[(y - y_C)/(x - x_C)\right]$, which value determines the number of the segment, say j, within the point lies. The unknown in-plane part of the normal vector is perpendicular to the j-th segment:

$$\begin{aligned} N_{x,j} &= \left(y_{j+1} - y_j\right)\sqrt{\frac{1 - N_{z,j}^2}{\left(y_{j+1} - y_j\right)^2 + \left(x_{j+1} - x_j\right)^2}} \\ N_{y,j} &= -\left(x_{j+1} - x_j\right)\sqrt{\frac{1 - N_{z,j}^2}{\left(y_{j+1} - y_j\right)^2 + \left(x_{j+1} - x_j\right)^2}} \end{aligned} \tag{7.95}$$

The expression in the square root comes from the normalization of $\vec{N}$.

The value of the permittivity depends on whether the point (x,y) lies inside or outside the polygon. The calculations of $\varepsilon(x,y)$ and $1/\varepsilon$ are made simultaneously with the normal vector calculus. After the angular segment in which the point (x,y) of the mesh in grating cell is determined (say the j-th one, as in eq.(7.95)), we can find the length of ρ_S between the central point and the polygon segment, shown in Fig.7.8. For this sake we show in Fig.7.9 the enlarged segment:

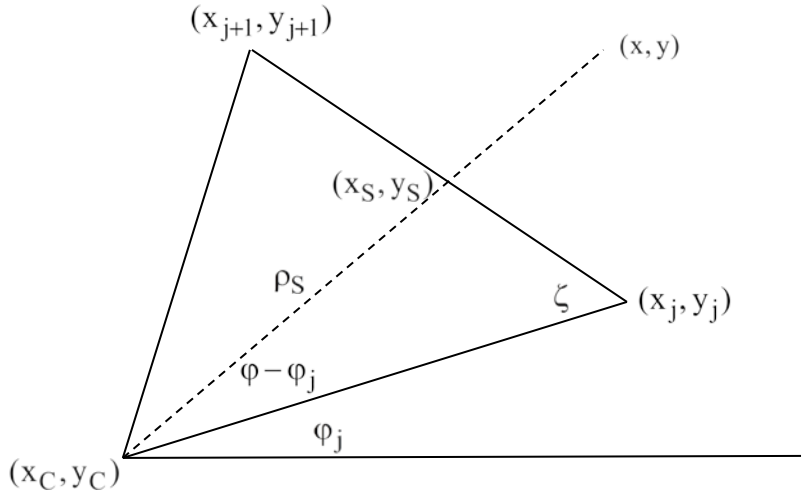


Fig.7.9. The j-segment of Fig.7.8 together with notations

The sine theorem gives the possibility to determine the angle ζ:

$$\frac{\sqrt{\left(x_{j+1}-x_j\right)^2+\left(y_{j+1}-y_j\right)^2}}{\sin(\varphi_{j+1}-\varphi_j)}=\frac{\sqrt{\left(x_{j+1}-x_C\right)^2+\left(y_{j+1}-y_C\right)^2}}{\sin(\zeta)} \tag{7.96}$$

wherefrom the radius ρ_S is given as:

$$\rho_S=\sin(\zeta)\frac{\sqrt{\left(x_C-x_j\right)^2+\left(y_C-y_j\right)^2}}{\sin(\pi-\zeta-\varphi+\varphi_j)} \tag{7.97}$$

Eq.(7.92) enables us to obtain the values of the permittivity (and its inverse):

$$\begin{aligned}&\rho(\varphi)<\rho_S(\varphi), \quad \varepsilon(x,y)=\varepsilon_{inside}\\ &\rho(\varphi)\geq\rho_S(\varphi), \quad \varepsilon(x,y)=\varepsilon_{outside}\end{aligned} \tag{7.98}$$

with $\rho=\sqrt{(x-x_C)^2+(y-y_C)^2}$.

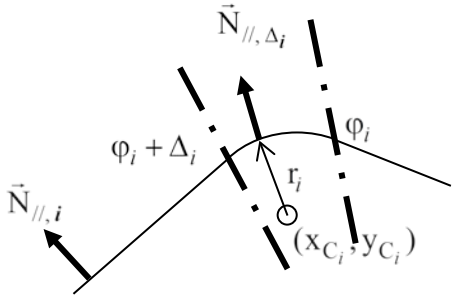


Fig.7.10. Schematic presentation of corner rounding

Concerning the edges, in reality the surfaces never have such, as etching always ends by rounding the corners, as shown in Fig.7.10. Let us consider that the rounding between the segments numbered *i*-1 and *i* is made preserving the values of N_z, and that in the cross-plane z = const., the rounding can be considered as circular, having a center in (x_{C_i},y_{C_i}) and radius r_i .The in-plane component of the normal vector follows the curvature radius and thus is given by equations, similar to eq.(7.95):

$$\begin{aligned}N_{x,\Delta_i}&=\left(x-x_{C_i}\right)\sqrt{\frac{1-N_{z,i}^2}{\left(y-y_{C_i}\right)^2+\left(x-x_{C_i}\right)^2}}\\ N_{y,\Delta_i}&=\left(y-y_{C_i}\right)\sqrt{\frac{1-N_{z,i}^2}{\left(y-y_{C_i}\right)^2+\left(x-x_{C_i}\right)^2}}\end{aligned} \tag{7.99}$$

The prolongation is more complicated, if the consecutive values of N_z at the two sides of the rounded corner differ significantly. In that case a linear interpolation of N_z between φ_i and $\varphi_i + \Delta_i$ can be applied.

7.6.2.3. Mutlivalued cross-sections

If the cross-section cannot be represented as a radial function, another possibility arises if it is a piecewise analytical function in x (or y), as shown in Fig.7.11, where we can use two different functions of x. We assume again that N_z is known, as it happens for z-independent profiles, for which it is simply null. In the upper part of the figure, for each value of x it is possible to calculate the normal vector on the profile:

$$\vec{N}_1 = \frac{\left(-f_1'(x), 1, N_z\right)}{\sqrt{1 + f_1'^2(x) + N_z^2}} \tag{7.100}$$

We can take this value to be the same for each y in the upper region A_1, so that the numerical Fourier transform is made only once in A_1 and once in A_2.

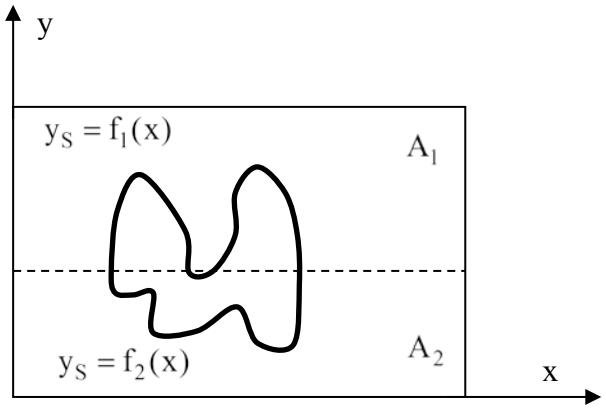


Fig.7.11. Piecewise analytical c ross-section of the grating surface at z = const.

The permittivity has to be calculated at each (x,y) mesh point:

$$\begin{aligned} y < y_S, \quad & \varepsilon(x,y) = \varepsilon_{\text{inside}} \\ y \geq y_S, \quad & \varepsilon(x,y) = \varepsilon_{\text{outside}}. \end{aligned} \tag{7.101}$$

7.6.4. Objects with cylindrical symmetry

Many periodic systems consist of inclusion having rotational cylindrical symmetry, like spheres, vertical cylinders, or ellipsoids with axe of rotation parallel to the z-axis, but also smooth surfaces, as presented in Fig.7.12.

These structures are characterized by a circular cross-section of the surface with the horizontal planes at z = const., but also with an independence on x and y of the values of N_z on each horizontal plane. In addition, due to the circular cross-sections, the angular component $N_\varphi = 0$ everywhere. Once z is fixed, the variation of the interface in the vertical

direction fixes the value of N_z, for example through eq.(7.89), wherefrom the radial normal vector component $N_\rho = \sqrt{1 - N_z^2}$. For each pair of (x,y) then:

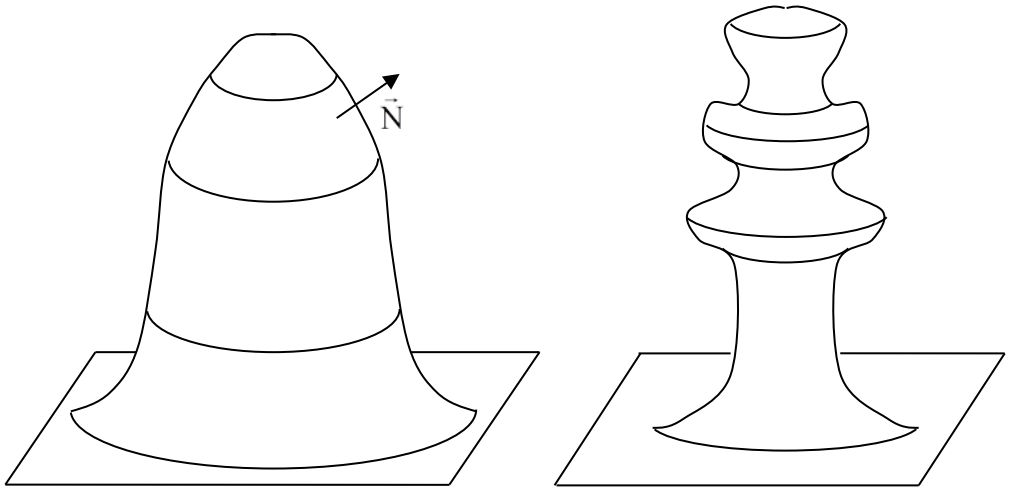


Fig.7.12. Several profiles with cylindrical symmetry.

$$N_x(x,y) = \frac{(x - x_C)N_\rho}{\sqrt{(x - x_C)^2 + (y - y_C)^2}}$$
$$N_y(x,y) = \frac{(y - y_C)N_\rho}{\sqrt{(x - x_C)^2 + (y - y_C)^2}} \qquad (7.102)$$

In addition, for profiles invariant in z-direction, $N_z = 0$ and $N_\rho = 1$.

The permittivity is given as a piecewise constant function:

$$\varepsilon(x,y) = \varepsilon_{inside}, \quad \text{if} \quad (x - x_C)^2 + (y - y_C)^2 < R^2(z),$$
$$\varepsilon(x,y) = \varepsilon_{outside}, \quad \text{if} \quad (x - x_C)^2 + (y - y_C)^2 \geq R^2(z) \qquad (7.103)$$

where R(z) is the radius of the profile surface for a given z.

Having obtained the values of the normal vector components and permittivity for each x,y enables us to calculate their Fourier transforms, either by Fast Fourier transform (FFT), or analytically, as shown in Sec.7.7.

7.6.5. Objects with elliptical cross-section

Similar simplification is possible for systems with elliptical cross-sections that have N_z = const. for z fixed. Such are the inclusions of vertical cylinders with elliptical cross-section, ellipsoids with one of the axes orientated in z-direction, but also all types of the structures shown schematically in Fig.7.12 that have elliptical or circular cross-sections.

Let us assume first that the ellipse axes are parallel to the x and y-axes. The more general case is discussed in Sec.7.7. The cross-section curve for z = const. is given by the equation:

$$\left(\frac{x_S - x_C}{a}\right)^2 + \left(\frac{y_S - y_C}{b}\right)^2 = R^2 \tag{7.104}$$

In order to obtain results similar to eq.(7.102), we introduce an elliptical coordinates, defined as:

$$\begin{aligned} \tilde{x} &= \frac{x - x_C}{a} \\ \tilde{y} &= \frac{y - y_C}{b} \end{aligned} \tag{7.105}$$

Using these notations, the ellipse becomes a circle, for which the considerations of the previous subsection apply. Thus

$$\begin{aligned} N_x(x,y) &= \frac{\frac{x - x_C}{a} N_\rho}{\sqrt{\left(\frac{x - x_C}{a}\right)^2 + \left(\frac{y - y_C}{b}\right)^2}} \\ N_y(x,y) &= \frac{\frac{y - y_C}{b} N_\rho}{\sqrt{\left(\frac{x - x_C}{a}\right)^2 + \left(\frac{y - y_C}{b}\right)^2}} \end{aligned} \tag{7.106}$$

with $N_\rho = \sqrt{1 - N_z^2}$, which remains constant for each fixed z.

Concerning the permittivity, it is determined in the same way as in eq.(7.103) for circular profile:

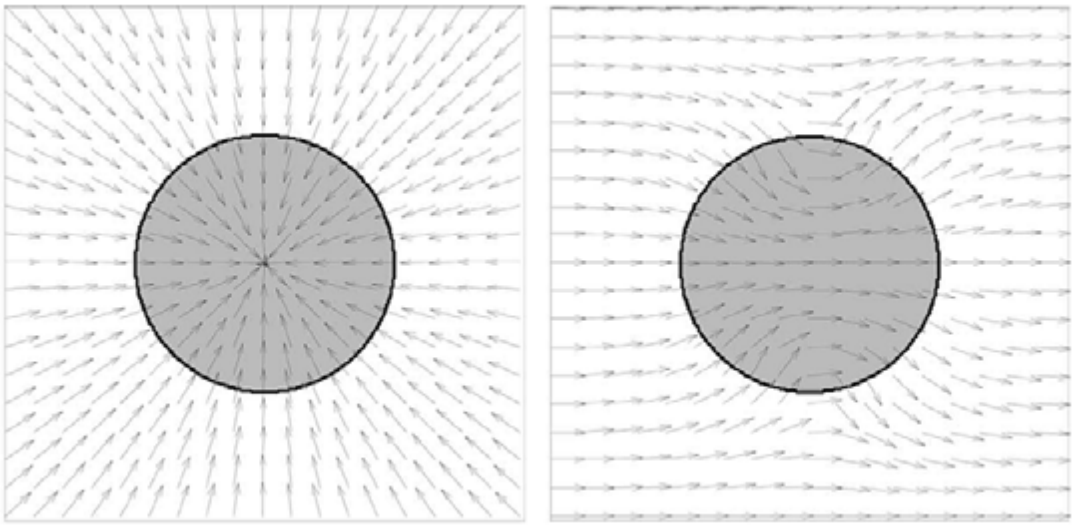
$$\begin{aligned} &\varepsilon(x,y) = \varepsilon_{inside}, \quad \text{if} \quad \left(\frac{x - x_C}{a}\right)^2 + \left(\frac{y - y_C}{b}\right)^2 < R^2(z), \\ &\varepsilon(x,y) = \varepsilon_{outside}, \quad \text{if} \quad \left(\frac{x - x_C}{a}\right)^2 + \left(\frac{y - y_C}{b}\right)^2 \geq R^2(z) \end{aligned} \tag{7.107}$$

with R(z) given by eq.(7.104). Sec.7.7 shows how to avoid the numerical FFT.

Remark on the prolongation of the normal vector

Special attention has recently been paid to the numerical implementation of the differential method for gratings having 2D periodicity formed by vertical holes or bumps that are invariant in z, and that have arbitrary cross-section in the xOy plane. A detailed study in the case of z-invariant geometry that applies for an eigenvalue/eigenvector technique of integration (FM or RCW method, see Chapter 13) is given in ref.[7.20], followed by several other works [7.21, 7.22]. It is necessary to note that the technique of prolongation of the normal vector as discussed in [7.20] can be applied also for z-dependent profiles with similar cross section; the difference is the renormalization factor $\sqrt{1 - N_z^2}$ for each z.

The authors compare several different formulations of the Fourier Modal method applied to structures with rectangular, circular, or elliptical cross-sections. These formulations include the classical formulation of Moharam and Gaylor that uses only the direct factorization rule, the formulation given by Lifeng Li [7.23, 7.31] that introduces two different Fourier transforms of the permittivity ε, namely $\lfloor\lceil\varepsilon\rceil\rfloor$ and $\lceil\lfloor\varepsilon\rfloor\rceil$, which are calculated by applying at first the inverse rule along one of the coordinates, and then the direct rule along the other one. This second formulation was made for rectangular and parallelogram cross-sections. For circular or elliptical (or other smooth) forms, it introduces a stepwise treatment of the profile, which appears more slowly convergent than the special techniques developed after.



(a) (b)

Fig.7.13. Two different prolongations of the normal vector for a circular inclusion. (a) Radial prolongation. (b) Electrostatic continuation of the normal vector for a circular cross-section inclusion inside a square grating cell (after [7.20]).

The third approach to the problem requires a prolongation of the normal vector (NV) to the profile within the entire grating cell. As already stressed, there are several possibilities to make this. A typical example is the radial prolongation, Fig.7.13a, which has been discussed in Sec.7.6.2.1 and 7.6.4 and it includes discontinuities of the normal vector on the cell boundaries, where the permittivity is continuous. Another approach proposed in [7.20] is the electrostatic one, which insures the continuity all over the cell and on its boundaries, except for on single points inside, Fig.7.13b.

Fig.7.14 shows the convergence rates for the transmitted zeroth order of a grating consisting of dielectric cylindrical inclusions with a circular cross-section with refractive index n = 1.5, in normal incidence from the substrate. The grating period is 2λ, the width of the grooves is λ, and the grating depth is $\lambda/(2n-1)$. The graph presents the diffraction efficiency in transmission as a function of the truncation order N using the three considered formulations: Moharam's original formulation, Li's formulation, and the formulation using the normal vector (NV) field. As usual the Fourier series run from – N to N, which yields 2N+1 Fourier coefficients for each of the two directions of periodic continuation, or $(2N+1)^2$ coefficients in total. As can be expected, both the original approach and the formulation by Li have worse convergence than when correctly taking into account the factorization rules for the tangential electric field and normal displacement components to the profile, where the permittivity is discontinuous [7.20]. It is necessary to stress that the difference in the

convergence rates is even more pronounced for metallic inclusions, having much larger optical contrast.

The fact that the discontinuity of the NV at the grating cell boundaries does not have observable influence on the convergence rates (provided that the NV is continuous at the discontinuities of the permittivity) is used later in sec.7.7, and enables us to replace the numerical Fourier transformation (made usually by the FFT programs) by analytical formulae using the cylindrical Bessel functions. This possibility could save computation time and avoid the discretization of the profile, when mapping it on the rectangular mesh necessary for the FFT.

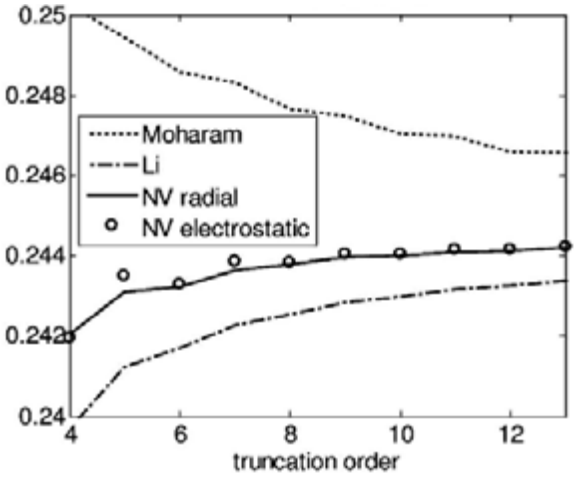
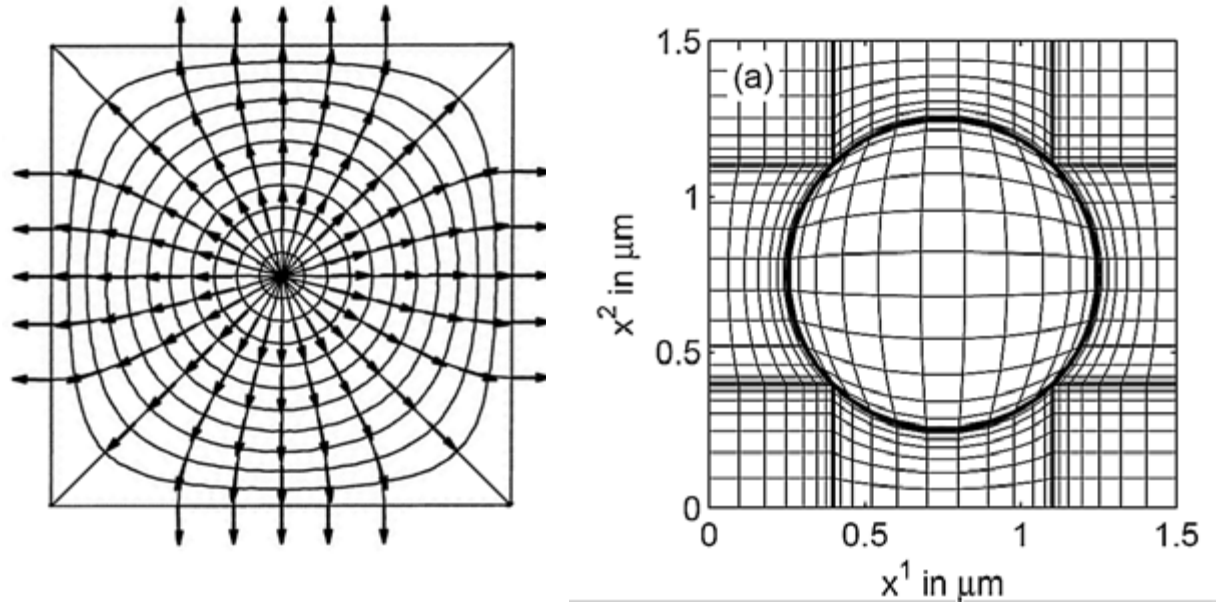


Fig.7.14. Convergence rates with respect to the truncation of the Fourier series for four different approaches used to model the diffraction by a cylindrical inclusion with circular cross-section.

Recently, Weiss et al. [7.24] proposed another alternative approach to treating smooth inclusions, by changing the coordinate system, so that its planes are parallel to the profile *and* to the grating sell walls (see Fig.7.15). If the transformed system is orthonormal, its coordinate lines are automatically tangential or perpendicular to the physical walls. If not, the Maxwell equations have to be rewritten in covariant vector form using the covariant and contravariant vector components.



(a) (b)

Fig.7.15. Coordinate lines and surfaces according to (a) [7.20] and (b) [7.24]

This approach is somehow equivalent to the normal vector prolongation, due to two main reasons:

(i) The NV approach defines in an unambiguous manner the normal vectors on the profile, giving a liberty to continue them all over the cell. The coordinate transformation is also defined on the profile and the outside boundaries, but can be chosen in different ways around the grating cell.

(ii) The change of the coordinate system introduces in the Maxwell equations the metric tensor $\mathbb{G}$ that multiplies the electric displacement and magnetic induction in the right-hand side of eq.(7.6), so that for the electric field we obtain the substitution:

$$\vec{D} = \varepsilon\vec{E} \quad \rightarrow \quad \mathbb{G}\varepsilon\vec{E}. \tag{7.108}$$

The normal vector approach acts in a similar manner by introducing the matrix Q_ε, given in eq.7.21, which makes the following substitution in the Fourier space, eq.(7.20):

$$\left[\vec{D}\right] = \left[\varepsilon\vec{E}\right] \quad \rightarrow \quad Q_\varepsilon\left[\vec{E}\right]. \tag{7.109}$$

7.6.6. Multiprofile surfaces

A grating with multiple bumps (or inclusions) inside the single cell could be treated by separation the cell into sub-cells, not necessarily rectangular, containing a single inclusion, as shown in Fig.7.16, where a specific cross-section at z = const. is separated into three regions A, B, and C. As far as the Fourier components of the normal vector, of the permeability and the permittivity have to be calculated for each value of z (if they depend on z), the separation into subcells can vary with z.

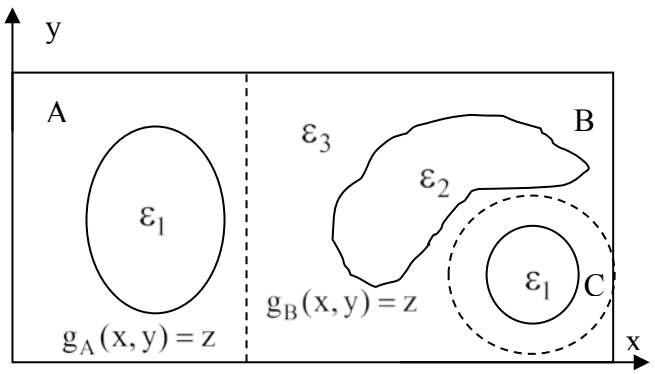


Fig.7.16. Cross-section at z = const. of a grating having different inclusions. The three different regions to be treated independently are separated by dashed lines.

The case schematized in Fig.7.17a can result from a surface covered by a thin layer of another substance, a layer that cannot be treated using eq.(7.87). The simplest possibility is to have different continuation of the normal vector inside each region. At first, the angle $\varphi = \arctan\left[(y - y_C)/(x - x_C)\right]$ for the point with coordinates (x, y) is calculated, and it is necessary to determine to which region the point belong. If it lies inside the innermost region C, the values of $\vec{N}(x,y) = \vec{N}_1(\varphi)$, where $\vec{N}_1(\varphi)$ is determined using one of the procedures discussed above for a single interface that is defined by the inner profile function.

If the point (x, y) lies in the outermost region, we take $\vec{N}(x,y) = \vec{N}_2(\varphi)$, where $\vec{N}_2(\varphi)$ corresponds to the second interface. In-between, we have two possibilities. The first choice is to divide the region into two subregions as indicated in Fig.7.17a with the dashed line. In each of them, $\vec{N}(x,y)$ is taken to be equal to its values on the adjacent profile, so that it is continuous everywhere where the permittivity and/or permeability are discontinuous.

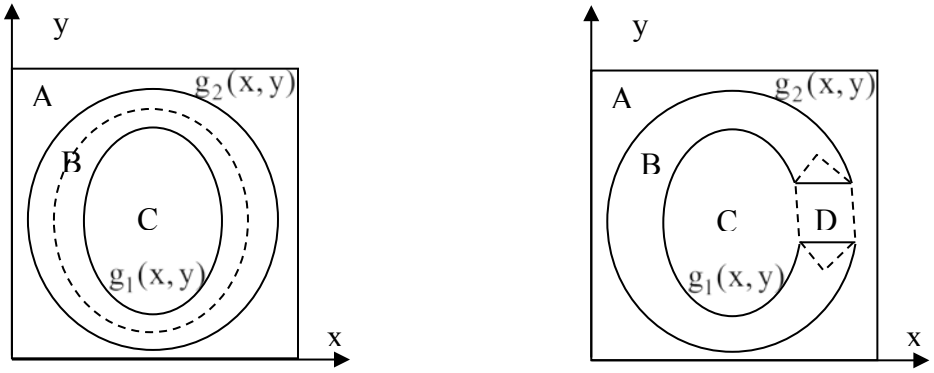


Fig.7.17. Structures with interpenetrating cross-section profiles

The second possibility is to introduce a linear interpolation inside region B, but it is necessary to know the distances ρ_1 and ρ_2 between the central point and the profiles along the ray with $\varphi = \arctan\left[(y-y_C)/(x-x_C)\right]$ fixed. Then:

$$\vec{N}_B(x,y) = \vec{N}_1(\varphi) + \left[\vec{N}_2(\varphi) - \vec{N}_1(\varphi)\right]\frac{\rho-\rho_1}{\rho_2-\rho_1}, \qquad (7.110)$$

$$\text{with} \quad \rho^2 = (x-x_C)^2 + (y-y_C)^2$$

Another specific case that appears in the studies of magnetic resonators is presented in Fig.7.17b. In can be treated in the same way is for the case in Fig.7.17a, but it is necessary to introduce a separate region D indicated in the figure and containing the opening, for which $\vec{N} = (0, 1, 0)$, for example.

7.7. Some cases of analytical Fourier transforms

There exist few but important cases when the Fourier transformation of the normal vector, of ε and/or of μ can be done analytically, as proposed in eq.(7.45) for grating with one-dimensional periodicity. In order to obtain this possibility, it is necessary that the z-component of the normal vector is invariant with respect to x and y for z fixed, and that the cross-section of the surface along the horizontal plane is circular, elliptic, rectangular, or rhombic. This concerns the examples with geometry given in Fig.7.1, Fig.7.6, and Fig.7.12. In particular, this procedure has been applied for the z-invariant structures, for which the Fourier modal method is more suitable.

We shall compare the convergence rates and the computation times of this analytical approach with the numerical FFT. Let us stress that the analytical determination of the Fourier components of the permittivity (and permeability for magnetic media) can be made for

elliptical (and circular) cross-sections, independently on the prolongation of the normal vector.

For instance, the analytical determination of the Fourier coefficients for 2D gratins has been implemented numerically for z- independent structures having circular cross sections, that are treated using the Fourier modal method. However, the Fourier transformation of permittivity (and permeability for magnetic materials) and of the normal vector tensor are made in the same manner for z-invariant profiles, as well as for objects that have geometry that is not vertically independent, such as spherical or elliptical inclusions, or profiles shown in Fig.7.12. At each value of z, it is necessary to recalculate the elements of the M-matrix.

Let us consider that for a value of z fixed, the normal vector z-component does not depend on x and y (but it can vary with z). We can use eq.(7.106) in the numerical FFT, but if the profile of the z-cross section has the form given by eq.(7.104), it is possible to avoid numerical Fourier transform. Let us consider the more general case with the ellipse axes that are *not* parallel to the unit cell walls, with an angle ψ between them, Fig.7.18. The ellipse equation in the initial coordinate system linked to the unit cell is given as:

$$\tilde{\rho} = R\,, \tag{7.111}$$

We have introduced new coordinates in order to obtain the canonical form (Eq.(7.111)) of the ellipse equation:

$$\begin{aligned} \tilde{x} &= \frac{x - x_c}{a}\cos\psi + \frac{y - y_c}{b}\sin\psi \\ \tilde{y} &= -\frac{x - x_c}{a}\sin\psi + \frac{y - y_c}{b}\cos\psi \\ \tilde{\rho} &= \sqrt{\tilde{x}^2 + \tilde{y}^2} \\ \mathrm{tg}\tilde{\varphi} &= \frac{\tilde{y}}{\tilde{x}} \end{aligned}\,, \tag{7.112}$$

The normal vector has components:

$$\begin{aligned} N_{\tilde{\rho}} &= \sqrt{1 - N_z^2} \\ N_{\tilde{\varphi}} &= 0 \end{aligned}\,, \tag{7.113}$$

that do not depend on $\tilde{\rho}$ and $\tilde{\varphi}$. Thus the Fourier transforms can be done analytically by considering separately the regions outside the ellipse A + B and inside it C (Fig.7.18). The second ellipse that is "concentric" to the profile is introduced in order to calculate the normal vector Fourier transforms.

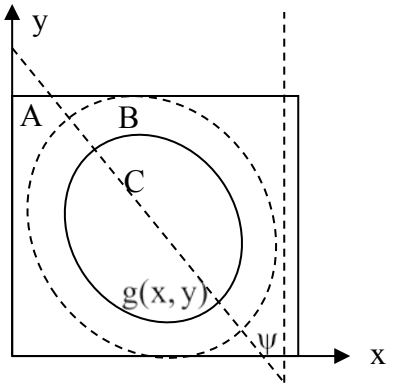


Fig.7.18. An inclined ellipsoidal profile g(x,y) with different permittivities in the regions A + B and C, together with a concentric ellipse (dash line) necessary for the normal vector transformation.

The Fourier transformation of the permittivity is done using the integral:

$$\varepsilon_{m,n} = \frac{\varepsilon_C}{d_x d_y}\int_C e^{-imK_x x - inK_y y}dxdy + \frac{\varepsilon_A}{d_x d_y}\int_{A+B} e^{-imK_x x - inK_y y}dxdy = \tag{7.114}$$

$$\frac{(\varepsilon_C - \varepsilon_A)}{d_x d_y} e^{-imK_x x_c - inK_y y_c}\int_0^R \tilde{\rho}\, d\tilde{\rho}\int_0^{2\pi} d\tilde{\varphi} e^{-imK_x a\tilde{\rho}\cos(\tilde{\varphi}+\psi) - inK_y b\tilde{\rho}\sin(\tilde{\varphi}+\psi)} + \frac{\varepsilon_A}{d_x d_y}\int_{A+B+C} e^{-imK_x x - inK_y y}dxdy$$

The second integral on the second line of eq.(7.114) is equal to the Kronecker's symbol:

$$\int_{A+B+C} e^{-imK_x x - inK_y y}dxdy = d_x d_y \delta_{m,0}\delta_{n,0} \tag{7.115}$$

while the first integral is evaluated analytically, after introducing two substitutions, defined in eq.(7.117):

$$\begin{aligned}&\int_0^R \tilde{\rho}\, d\tilde{\rho}\int_0^{2\pi} d\tilde{\varphi} e^{-imK_x a\tilde{\rho}\cos(\tilde{\varphi}+\psi) - inK_y b\tilde{\rho}\sin(\tilde{\varphi}+\psi)} = \int_0^R \tilde{\rho}\, d\tilde{\rho}\int_0^{2\pi} d\tilde{\varphi} e^{-i\tilde{\rho}K_{mn}\sin(\tilde{\varphi}+\psi+\chi_{mn})} \\ &= 2\pi\int_0^R \tilde{\rho}\, d\tilde{\rho}\, J_0(\tilde{\rho}K_{mn}) = \frac{2\pi R}{K_{mn}} J_1(RK_{mn})\end{aligned} \tag{7.116}$$

where J_p is the p^{th} order Bessel function, and we have used the substitutions:

$$\begin{aligned}K_{mn} &= \sqrt{m^2K_x^2a^2 + n^2K_y^2b^2} \\ tg\chi_{mn} &= \frac{mK_x a}{nK_y b}\end{aligned} \tag{7.117}$$

Similar integrals are obtained for the Fourier transforms of the normal vector components, which must be continuous across the interface g(x,y). As shown in Fig.7.18, we introduce another ellipse $g_2(x,y)$, that obeys an equation, similar to eq.(7.111), but having larger dimensions, $\tilde{\rho} = R_2$. The evaluation of

$$\left(N_x^2\right)_{m,n} = \frac{1}{d_x d_y}\int_{A+B+C} N_x^2(x,y)e^{-imK_x x - inK_y y}dxdy \tag{7.118}$$

can be made in two different regions, A and B + C, taking different prolongations of the normal vector. The important feature in the application of the inverse and/or the direct factorization rules is that they require that the normal vector and the permittivity have no common points of discontinuity.

The vector normal to the surface is proportional to the gradient of the surface equation:

$$\begin{aligned}N_x(x,y) &\sim \frac{\partial}{\partial x}\left[\left(\frac{x-x_c}{a}\right)^2 + \left(\frac{y-y_c}{b}\right)^2 - R^2\right] = 2\frac{x-x_c}{a^2} = \frac{2}{a}\tilde{\rho}\cos(\tilde{\varphi}+\psi) \\ N_y(x,y) &\sim \frac{\partial}{\partial y}\left[\left(\frac{x-x_c}{a}\right)^2 + \left(\frac{y-y_c}{b}\right)^2 - R^2\right] = 2\frac{y-y_c}{b^2} = \frac{2}{b}\tilde{\rho}\sin(\tilde{\varphi}+\psi)\end{aligned} \tag{7.119}$$

After the normalization, we obtain that:

$$N_x(x,y) = b\sqrt{\frac{1-N_z^2}{a^2+b^2}}\cos(\tilde{\varphi}+\psi) \equiv \tilde{N}_x\cos(\tilde{\varphi}+\psi)$$
$$N_y(x,y) = a\sqrt{\frac{1-N_z^2}{a^2+b^2}}\sin(\tilde{\varphi}+\psi) \equiv \tilde{N}_y\sin(\tilde{\varphi}+\psi) \qquad (7.120)$$

inside the region B + C, i.e., continuous across the real profile g(x,y). As they can be discontinuous on the fictitious profile $g_2(x,y)$, we can take them in a form that allows for the Fourier transform in an analytical form, for example,

$$N_x(x,y) = \sqrt{1-N_z^2}$$
$$N_y(x,y) = 0 \qquad (7.121)$$

as shown in Fig.7.19, which illustrates the geometry of this approach in the prolongation of the normal vectors for a circular cross-section, and has to be compared with Fig.7.13a and Fig.7.15, which make either a radial prolongation of the normal vector that is discontinuous at the cell boundaries, or a smooth prolongation that is continuous everywhere in the cell.

As far as the normal vector components participate in pairs ($N_{x,y}^2$ and N_xN_y), its sign in a given direction plays no role and can be taken as most convenient.

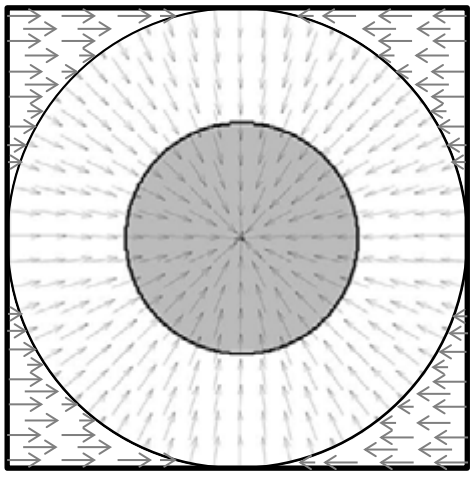


Fig.7.19. Piecewise prolongation of the normal vector, applied for a circular cylinder profile

Similar to eq.(7.114), we can write

$$\left(N_x^2\right)_{m,n} = \frac{\tilde{N}_x^2}{d_x d_y}\int\limits_{B+C}\cos^2(\tilde{\varphi}+\psi)\,e^{-imK_x x-inK_y y}dxdy+\frac{1-N_z^2}{d_x d_y}\int\limits_{A}e^{-imK_x x-inK_y y}dxdy$$
$$=\frac{\tilde{N}_x^2+N_z^2-1}{d_x d_y}\int\limits_{B+C}\cos^2(\tilde{\varphi}+\psi)\,e^{-imK_x x-inK_y y}dxdy+\frac{1-N_z^2}{d_x d_y}\int\limits_{A+B+C}e^{-imK_x x-inK_y y}dxdy \tag{7.122}$$

The second integral is taken over the entire unit cell, and it is given by eq.(7.115). The first integral is developed as follows:

$$\int\limits_{B+C}\cos^2(\tilde{\varphi}+\psi)\,e^{-imK_x x-inK_y y}dxdy =$$
$$e^{-imK_x x_c-inK_y y_c}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\int\limits_0^{2\pi}d\tilde{\varphi}\frac{1}{4}\left[2+e^{2i(\tilde{\varphi}+\psi)}+e^{-2i(\tilde{\varphi}+\psi)}\right]e^{-imK_x a\tilde{\rho}\cos(\tilde{\varphi}+\psi)-inK_y b\tilde{\rho}\sin(\tilde{\varphi}+\psi)} \tag{7.123}$$

The first term in the square brackets can be expressed in the form given by eq.(7.116):

$$\frac{1}{2}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\int\limits_0^{2\pi}d\tilde{\varphi}e^{-imK_x a\tilde{\rho}\cos(\tilde{\varphi}+\psi)-inK_y b\tilde{\rho}\sin(\tilde{\varphi}+\psi)}=\frac{\pi R}{K_{mn}}J_1\left(RK_{mn}\right) \tag{7.124}$$

The second and the third terms in eq.(7.123) can also be expressed in the terms of Bessel functions:

$$\frac{1}{4}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\int\limits_0^{2\pi}d\tilde{\varphi}e^{\pm 2i(\tilde{\varphi}+\psi)-imK_x a\tilde{\rho}\cos(\tilde{\varphi}+\psi)-inK_y b\tilde{\rho}\sin(\tilde{\varphi}+\psi)}$$
$$=\frac{1}{4}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\int\limits_0^{2\pi}d\tilde{\varphi}e^{\pm 2i(\tilde{\varphi}+\psi)-i\tilde{\rho}K_{mn}\sin(\tilde{\varphi}+\psi+\chi_{mn})}$$
$$=\frac{1}{4}e^{\mp 2i\chi_{mn}}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\int\limits_0^{2\pi}d\tilde{\varphi}e^{\pm 2i(\tilde{\varphi}+\psi+\chi_{mn})-i\tilde{\rho}K_{mn}\sin(\tilde{\varphi}+\psi+\chi_{mn})} \tag{7.125}$$
$$=\frac{\pi}{2}e^{\mp 2i\chi_{mn}}\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\,J_2\left(\tilde{\rho}K_{mn}\right)$$

so that their sum is equal to:

$$\frac{\pi}{2}\left(e^{-2i\chi_{mn}}+e^{-i\chi_{mn}}\right)\int\limits_0^{R_2}\tilde{\rho}\,d\tilde{\rho}\,J_2\left(\tilde{\rho}K_{mn}\right)=$$
$$\frac{\pi}{K_{mn}^2}\cos(2\chi_{mn})\left[2-2J_0\left(R_2K_{mn}\right)+R_2K_{mn}J_2\left(R_2K_{mn}\right)\right] \tag{7.126}$$

The Fourier transformation of N_y^2 can be performed in exactly the same manner, but it is more direct to use the relation:

$$\left(N_y^2\right)_{m,n}=\sqrt{1-N_z^2}\left[\delta_{m,0}\delta_{n,0}-\left(N_y^2\right)_{m,n}\right] \tag{7.127}$$

The evaluation of the mixed term contains the product of $\sin(\tilde{\varphi}+\psi)\cos(\tilde{\varphi}+\psi) = \frac{1}{2}\sin 2(\tilde{\varphi}+\psi)$. Thus $\left(N_x N_y\right)_{m,n}$ result in the same form as eqs.(7.125) and (7.126), by replacing $\cos(2\chi_{mn})$ with $\sin(2\chi_{mn})$.

It is necessary to stress out the usefulness of these analytical results, when compared with the Fast Fourier transform. First, FFT need discretization that for smooth but not rectangular profile introduces some type of stepwise approximation. This stepwise approximation requires greater number of discretizations and thus longer computation times, which can be a significant disadvantage in the case of 2D periodicity. Second, when the profile is z-dependent, the evaluation of the Fourier transforms has to be made on each integration step, and this additionally worsens the computation time problem.

7.8. Integrating schemes

Numerical solution of a system of ordinary differential equations is a mature domain due to the enormous amount of physical and technical applications. Unfortunately, the grating problem represents one of the worst tasks for the theory of ordinary differential equations, because the system to be integrated is a *stiff* one. To better understand the problem, let us consider the case of a homogeneous layer that introduces no coupling between the diffraction orders. The solution of the diffraction problem contains waves propagating up- and downwards (in z-direction). These are plane waves, propagating or evanescent inside the layer. In lossless medium, their constant of propagation in z can be real, or imaginary, depending on the number of diffraction order under consideration:

$$\gamma_m = \pm\sqrt{(k_0 n)^2 - \alpha_m^2}\,, \tag{7.128}$$

with

$$\alpha_m = \alpha_0 + mK\,. \tag{7.129}$$

The real values of γ are bounded by $k_0 n$, but the imaginary parts are not bounded, as their asymptotic values for large |m| are given by:

$$\mathrm{Im}(\gamma_m) = \pm\,|m|\,K\,. \tag{7.130}$$

From the point of view of the theory of ordinary differential equations this means that the eigenvalues of the system differ significantly in magnitude, i.e., the differential system is *stiff*. The greater the difference, the more unstable the solution. On the other hand, the solution of the diffraction problem requires sufficient number of Fourier components of the profile function and electromagnetic field to be correctly represented by the truncated Fourier series, thus the necessity to work with large number of Fourier components, and thus the increasingly greater the stiffness of the differential system, i.e., more instable the solution with respect to the length and number of integration steps. The theory concludes that the so called *explicit* integration schemes are most instable for such problem, whatever their order, and *implicit* methods have to be used. The problem with the implicit methods is that they need one matrix inversion and several more matrix operations on each integration step, when compared with the explicit methods, so that the choice is not evident to ensure the most efficient integration scheme.

Let us recall the basic principle of the first-order explicit and implicit schemes. In a first-order approximation, the solution of the differential system:

$$\frac{d}{dz}F(z) = M(z)F(z). \tag{7.131}$$

between two consecutive points $z = z_j$ and $z = z_{j+1}$ can be searched in developing in series:

$$F(z_{j+1}) = F(z_j) + (z_{j+1} - z_j)M(z)F(z). \tag{7.132}$$

If $M(z)$ and $F(z)$ are evaluated in $z = z_j$, this leads to the first-order explicit integration (Euler's) scheme:

$$F(z_{j+1}) = \left[\mathbb{I} + hM(z_j)\right]F(z_j). \tag{7.133}$$

where $\mathbb{I}$ is the unit matrix, and $h = (z_{j+1} - z_j)$.

If $M(z)$ and $F(z)$ are evaluated in $z = z_{j+1}$, we obtain the first-order implicit (inverted or backward Euler's) scheme:

$$F(z_{j+1}) = \left[\mathbb{I} - hM(z_{j+1})\right]^{-1} F(z_j). \tag{7.134}$$

The theory says that this scheme is more stable, but it needs one matrix inversion on each step. A combination of the two must provide even better results, because it uses a half of the previous step:

$$F(z_{j+1}) = \left[\mathbb{I} - \frac{1}{2}hM(z_{j+1})\right]^{-1}\left[\mathbb{I} + \frac{1}{2}hM(z_j)\right]F(z_j). \tag{7.135}$$

However, we need one additional matrix multiplication. In what follows we use these two single-point first-order methods under the names **Expl 1** (single point explicit Euler integration) and **Impl 1**, eq.(7.135) and compare the convergence with respect to the total number of integration points with several other more sophisticated integration schemes for two different metallic gratings in TM polarization, the most difficult combination when using the differential method.

The advantage of these formulations is that they all are single-step ones, and do not need a storage of the intermediate results on several integration steps. They can be easily programmed and don't need additional memory storage at each step. However, if we refer to one of the most relevant sources [7.25], we see that "*this is the generic disease of stiff equations: We are required to follow the variation in the solution on the shortest length scale to maintain the stability of the integration, even though accuracy requirements allow a much larger stepsize*." This means that *a priori* choice of the integration step without adaptive control and change in the step length cannot produce stable and relevant results. Unfortunately, it is quite difficult to use adaptive-step methods, because they require much longer computation times, as it is necessary to repeat the integration process several times when changing the integration step length. This is why we concentrate our attention to fixed-step algorithms.

Fixed-step *multistep explicit* methods have been used from decades in the differential method programming. The best results have been obtained when combined with an implicit correction by using a predictor-corrector scheme, as described further on. However, referring again to [7.25], "*high order* does not always mean *high accuracy*." It will be more useful, if larger integration step is obtained with high order or multistep methods, which is not obvious, as we observe on several numerical examples.

We have used three simple integration schemes, the single-point implicit or explicit scheme, as well as a 4-point predictor-corrector method (**PCM 4**). It contains two steps, the first one representing an Adams-Bashforth explicit 4-point scheme [7.26], described by the equation

$$F(z_{j+5}) = F(z_{j+4}) + h\left[\frac{1901}{720}F'(z_{j+4}) - \frac{1387}{360}F'(z_{j+3}) + \frac{108}{30}F'(z_{j+2}) - \frac{637}{360}F'(z_{j+1}) + \frac{251}{720}F'(z_j)\right]. \quad (7.136)$$

with

$$F'(z_j) = M(z_j)F(z_j). \quad (7.137)$$

The corrector step is a 4-point Adams-Moulton integration:

$$F(z_{j+4}) = F(z_{j+3}) + h\left[\frac{251}{720}F'(z_{j+4}) + \frac{646}{720}F'(z_{j+3}) - \frac{264}{720}F'(z_{j+2}) + \frac{106}{360}F'(z_{j+1}) - \frac{19}{720}F'(z_j)\right], \quad (7.138)$$

which is an implicit scheme [7.27]. However, contrary to the other explicit schemes (BDF), it does not require inverting a matrix, it just makes one step back as a corrector.

An extension of eq.(7.134) to a multistep algorithm results in multistep implicit method, called also backward differentiation formulae (BDF) [7.28]. Typical 3-point and 5-point formulae take the form:

$$\text{BDF3}: \quad F(z_{j+3}) = \left[\mathbb{I} - hM(z_{j+3})\right]^{-1}\left[\frac{18}{11}F(z_{j+2}) - \frac{9}{11}F(z_{j+1}) + \frac{2}{11}F(z_j)\right]. \quad (7.139)$$

$$\text{BDF5}: \quad F(z_{j+5}) = \left[\mathbb{I} - hM(z_{j+5})\right]^{-1}\left[\frac{300}{137}F(z_{j+4}) - \frac{300}{137}F(z_{j+3}) + \frac{200}{137}F(z_{j+2}) - \frac{75}{137}F(z_{j+1}) + \frac{12}{137}F(z_j)\right]. \quad (7.140)$$

It is evident that BDF3 (called further on **Impl 3**) requires a starting method for the first two points, and BDF5 (**Impl 5**), for the first 4 points. The same is valid for PCM in eqs.(7.136) – (7.138).

The second-order Runge-Kutta method is given in the form:

$$\begin{aligned} k_1 &= hM(z_j)F(z_j) \\ k_2 &= hM(z_{j+1/2})\left[F(z_j) + \frac{1}{2}k_1\right] \\ F(z_{j+1}) &= F(z_j) + k_2 \end{aligned} \quad (7.141)$$

which has an error proportional to h^3. It is also called a midpoint point, because it requires the evaluation of the functions at the middle of the step, i.e., twice the number of the steps of the other tree methods discussed above.

The classical fourth-order Runge-Kutta method also uses midpoint values:

$$\begin{aligned} k_1 &= hM(z_j)F(z_j) \\ k_2 &= hM(z_{j+1/2})\left[F(z_j) + \frac{1}{2}k_1\right] \\ k_3 &= hM(z_{j+1/2})\left[F(z_j) + \frac{1}{2}k_2\right] \\ k_4 &= hM(z_{j+1/2})\left[F(z_j) + k_3\right] \\ F(z_{j+1}) &= F(z_j) + \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4 \end{aligned} \quad (7.142)$$

with an error of the order of h^5. These two methods are higher-order explicit methods that do not need matrix inversions during the integration. However, as already stressed, higher order does not mean larger steps.

The highest possible order in h can be obtained theoretically, by using the **eigentechnique**:

$$F(z_{j+1}) = V\left(e^{\gamma h}\right)V^{-1}F(z_j) \tag{7.143}$$

where V is a matrix containing the eigenvectors of M and the exponential term in the round brackets represents a diagonal matrix constructed using the eigenvalues γ of M. In practice, this method does not increase the stability, because it remains a single-point explicit method. Moreover, it requires much longer computation time because of the requirement to solve an eigenvalue/eigenvector problem at each integration step.

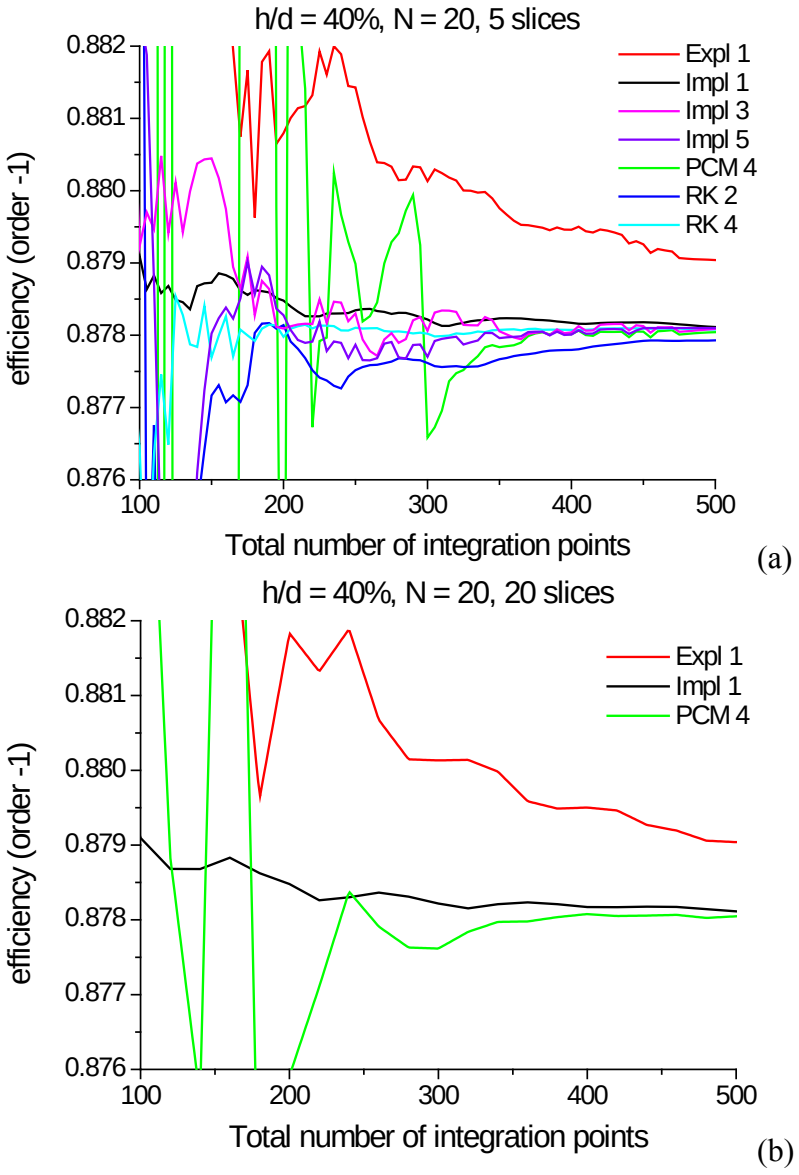


Fig.7.20. Aluminum grating with period 0.5 μm and depth 0.2 μm used in -1st order Littrow mount at 0.6328 μm wavelength in TM polarization. Convergence with respect to the total number of integration points, truncation to 41 Fourier harmonics and using 5 (a) and 20 (b) slices in the S-matrix algorithm. The acronyms for the methods are defined in the text.

The first example concerns a typical commercial sinusoidal aluminum grating that has very high efficiency in TM polarization. It supports a single diffracted order in -1st order Littrow mount and has a modulation depth-to-period ratio of 40%. Fig.7.20 presents a numeric test of the efficiency calculated for a different number of integration points using several integrations schemes. Due to the polarization and the grating material, it is necessary to separate the integration into several slices (5 in this case) in order to avoid numerical loss of precision, the results of the integration in two consecutive slices connected to each other by the use of the S-matrix algorithm. In Fig.7.20b we have presented a part of the results, obtained with 20 slices, instead of 5. The comparison between the two cases show that 5 slices are sufficient, the weaker oscillation for 20 slices are due mainly to the fact that the horizontal scale is less dense, because the step in the total number of integration points is an integer times the number of slices. The truncation parameter N = 20, i.e., totally 41 Fourier harmonics of the field are used in the calculations.

As can be concluded, an absolute precision within 1% is rapidly obtained whatever the method used, with the total number of points of the order of 200. However, the predictor corrector method is less stable when the number of points is smaller than 300. Implicit methods are more stable, as expected, and result in an error smaller than 0.1% even for the number of points less than 200. It is interesting to observe than the first-order implicit method is more stable than the higher implicit methods, probably because it contains a middle-point evaluation of the field derivative, as seen in eq.(7.135). It requires a little bit longer computation time than the other two implicit methods, because of the additional matrix multiplication. The explicit method, which is the fastest one, shows slower convergence, as expected, whereas the performance of the higher-order RK methods competes with the implicit methods.

Table 7.1 compares the computation times of the different methods for the two investigated cases (Fig.7.20 and the following Fig.7.21). For comparison, (null) indicates the time without any operation due to the integration, and that is necessary for the construction of the M-matrix and the use of the S-matrix propagation algorithm, as described in Appendix 7.A. The fastest method is the single-point explicit method, but as expected it is less precise given the same number of integration points, Fig.7.20. The implicit single-step middle-point method shows stability similar to the 4-th order Runge-Kutta method, but is slightly more rapid. The predictor-corrector method is less stable and requires longer computation times.

Table 7.1. Computation times of the different methods described in the text for the two cases with groove depth to groove period equal to 40% and 200%

Method	N = 20, slices 5 int. points 400, modulation 40%	N = 50, slices 35 int.points 1500, modulation 200%
Expl. 1	1.41 s	72.3 s
Impl.1	3.55 s	187.6 s
Impl.3	2.67 s	157.8 s
Impl.5	2.81 s	168.9 s
PCM 4	2.14 s	120.0 s
RK 2	2.70 s	144.9 s
RK 4	4.70 s	252.5 s
eigentechnique	7.54 s	360.0 s
(null)	0.68 s	38.5 s

It is necessary to stress out that in reality, the computation times are shorter than listed in the Table, because when the truncation N is smaller (usually 20 is sufficient), the number of slices for the S-matrix algorithm is smaller (due to the smaller number of evanescent orders taken into account); in addition, the total number of integration points used for constructing the Table are chosen to obtain 0.1% relative error, whereas in most of the cases just 1% is sufficient. The computation time grows linearly with the total number of integration points, as well as with the number of slices used in the S-matrix algorithm. The time dependence concerning the truncation N in the Fourier series grows as $N^3 - N^{3.5}$, because this parameter determines the size of the matrices.

When the total integration length is multiplied by 5, the number of integration points required is also multiplied by the same factor, as observed in Fig.7.21. A grating twice as deep as the period, acts almost like a flat mirror in TM polarization, with the efficiency in order -1 hardly exceeding 1%. Due to the large depth, the absorption is increased, so that the reflectivity in order 0 is equal to 56.78%. We compare the convergence in the weak -1st order, so that even a small absolute error appears as a large relative error that can be easily observed in the figures. The number of Fourier harmonics (truncation parameter 2N+1) also has to be increased by a factor of 2.5 to 101 (N = 50). The number of slices in the S-matrix algorithm is increased seventh-fold to 35.

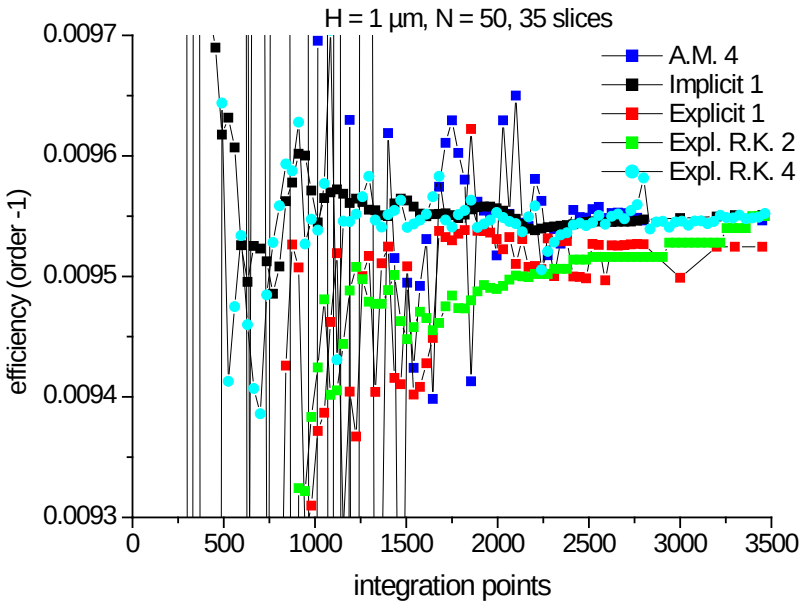


Fig.7.21. Aluminum grating with period 0.5 µm and depth 1 µm used in -1st order Littrow mount at 0.6328 µm wavelength in TM polarization. Convergence with respect to the total number of integration points, truncation to 101 Fourier harmonics and using 35 slices in the S-matrix algorithm.A.M.4 – forth-order Adams-Moulton scheme, Implicit 1 – single-point implicit scheme, Explicit 1 – single-point explicit scheme, Expl2.R.K.2 and 4 – explicit Runge-Kutta method of order 2 and4, respectively.

The first Fig.7.21 compares several explicit integration schemes with the single-point implicit method. The main conclusion to be drawn is that the best scheme remains the implicit method, only the explicit Runge-Kutta fourth-order scheme seems to compete in convergence rate with respect to the total number of integration points, but somehow slower.

The comparison of several implicit methods confirms the general idea that multistep choice does not necessarily improve the stability (Fig.7.22). When compared with Fig.7.21, the implicit methods are characterized by smaller oscillations when the number of steps is

increased, but the most rapid convergence is obtained with the simplest procedure, single-step method (let us remark again that we use the middle-point calculations, as in eq.(7.135)). Like all the other implicit methods, it requires a single matrix inversion on each integration step, but needs less memory storage, and avoids several matrix sums and multiplication by different constants, necessary for the multipoint methods.

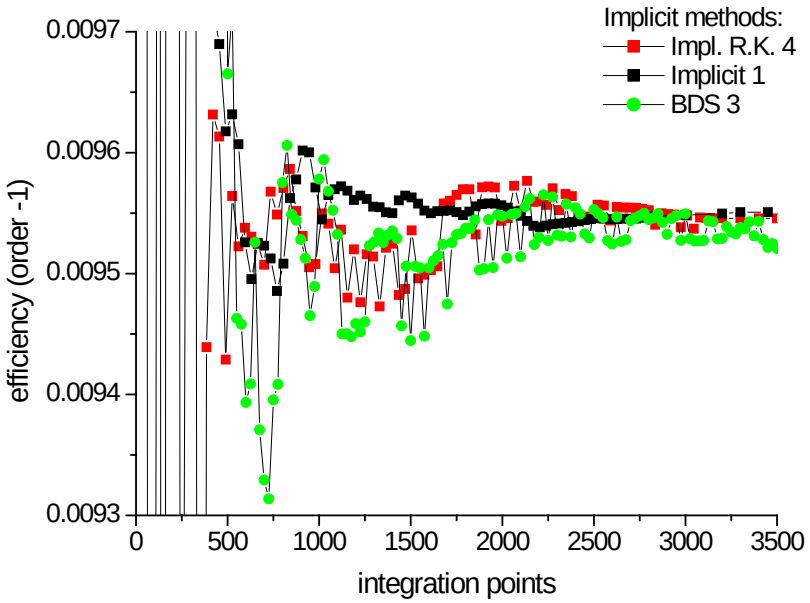


Fig.7.22. Same as in Fig.7.21 but for three implicit methods of different order.

7.9. Staircase approximation

As already discussed, if the surface interface is z-invariant (entirely or piecewisely), the integration of the system of ordinary differential equations along z can be done via eigenvalue/eigenvector technique, eq.(7.143), because the M-matrix containing the coefficients of the differential equations does not depend on z. The enormous interest in this approach can be explained by the simple technique of integration, much easier to understand and apply than the theory of numerical methods of integration of ordinary differential equations.

The idea is sketched in Fig.7.23, where a sinusoidal surface-relief grating is approximated by a 5-stairs profile. While this approximation (with sufficient number of steps M, depending on the groove depth) works quite well in TE polarization, the TM case presents a convergence rate with respect the truncation number of Fourier components of the field much slower than the ordinary differential method (no staircase approximation), see Fig.7.24. Moreover, the greater the number of vertical slices M, the greater the truncation number required.

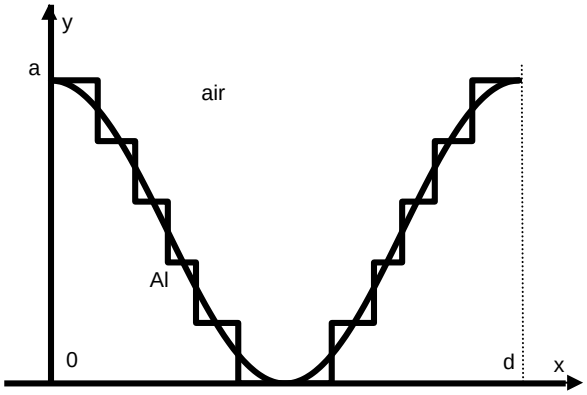


Fig.7.23. Schematical approximation of a sinusoidal grating profile, approximated by a 5-step staircase profile.(after [7.29]).

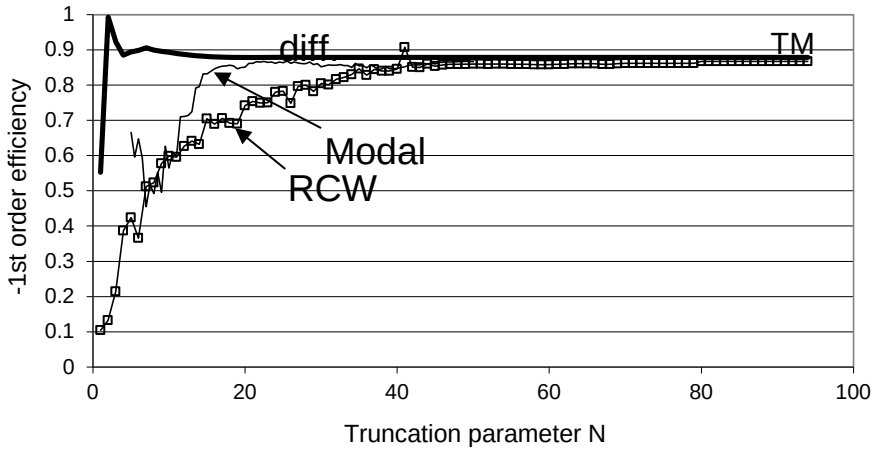


Fig.7.24. Convergence of the minus-first-order efficiency in TM polarization of the FMM (RCW) and the exact modal method (indicated on the figure) for a sinusoidal grating in a staircase presentation with $\tilde{M} = 20$*, as compared to the convergence of the differential method for a smooth sinusoidal profile (curve "diff."). Period d = 0.5 μm, groove depth a = 0.2 μm, aluminum refractive index* $n_{Al} = 1.3 + i7.6$*, illuminated at 40° incidence with wavelength* $\lambda = 0.6328$ *μm, (after [7.29)].*

A detailed analysis of this problem can be found in [7.14, 7.29], but the basic idea is quite simple. The staircase approximation substitutes the otherwise smooth sinusoidal profile by a profile that has sharp edges. The greater the number of stairs, the greater is the number of edges. It is well-known from general electromagnetism that edges introduce electric field singularities. While in TE polarization the only electric field components are tangential to the profile (in y-direction), thus have no discontinuities and singularities, this is not the case in TM polarization. This can be observed in Fig.7.25. At the edges of each step, a sharp maximum of the electric field is observed. These maxima are not a numerical artifact, they represent the physical effect of introducing edges to replace a smooth profile. These sharp variations of the field require larger number of Fourier components to be correctly represented. Moreover, the greater the number of slices (stairs), the greater the number of the maxima, thus the greater the truncation number required. Numerical experiment has shown that this phenomenon has nothing to do with the integration (eigenvalue/vector) technique, because the results of the convergence rate and field maps are the same for the staircase approximation when using the RCW technique or the differential method.

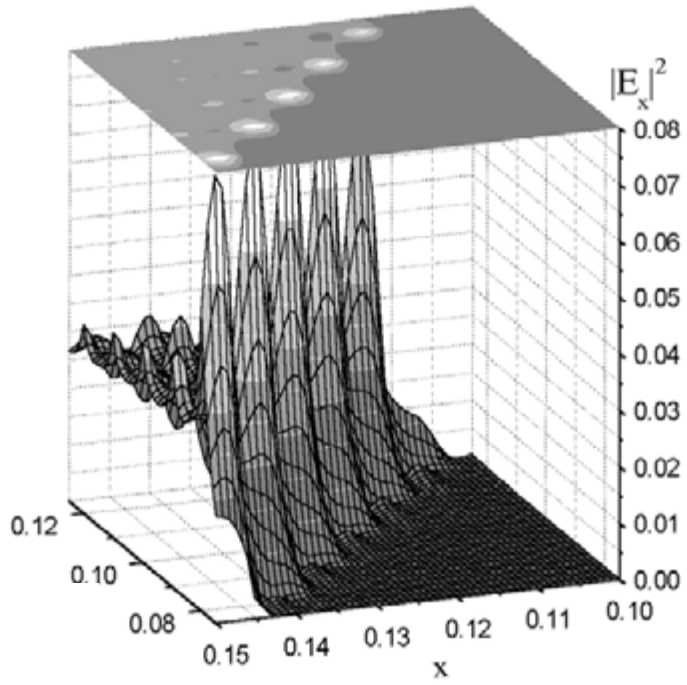


Fig.7.25. Spatial field distribution of $|E_x|^2$ in the vicinity of several steps inside a groove of a 10-step staircase profile, used to approximate the sinusoidal grating under study in TM polarization. The grating parameters are the same as in Fig.7.24, after [7.29].

On the contrary, if the true smooth profile is treated by the differential method by using a numerical integration of the ordinary differential system with the elements of the M-matrix depending on z, there is no such singularities of the electric field (Fig.7.26), so that the convergence with respect to the number of Fourier harmonics is much faster, provide the correct factorization rules are used (Fig.7.24).

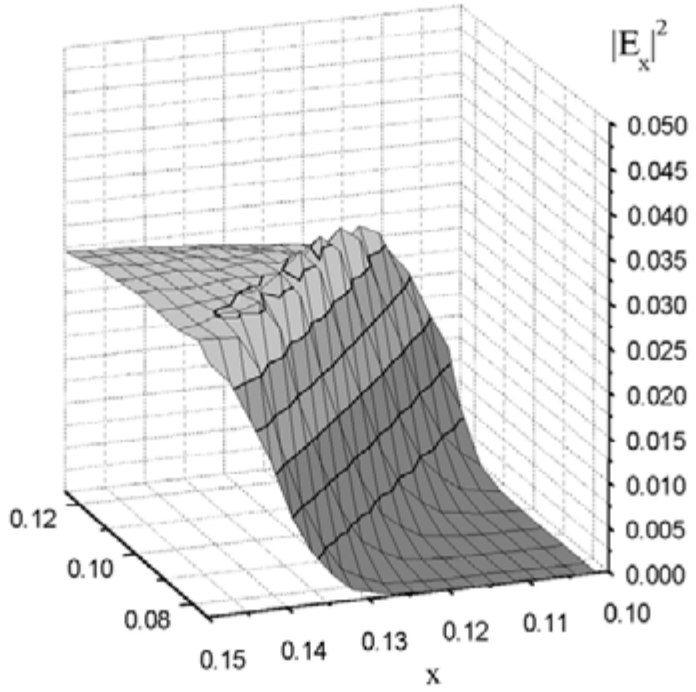


Fig.7.26. The same as in Fig.7.25 but calculated using the differential method, after [7.29].

Recently, some authors [7.30] have proposed to maintain the eigenvalue/vector technique, but to use the correctly determined Fourier presentation of the profile, i.e., the correct factorization rules, as presented in eqs. (7.39) – (7.44), instead of lamellar-profile factorization, eqs.(7.56) – (7.65), at each step. This is equivalent to using the formulation proposed by the differential method for a smooth profile, i.e. avoiding the field singularities at the edges, but to use the eigenvalue/vector technique of integration by assuming that the modified M-matrix, as given by eqs. (7.39) — (7.44), is z-invariant across each step height. We have already tried this in [7.29] and the conclusion was that using this approach, the number of steps (stairs) has to be relatively larger that by using some better adapted integration technique. And indeed, the eigenvalue/approach to a z-dependent system is equivalent to the rectangular rule with equidistant points of integration, one of the worst choices, as known from the theory of ordinary differential equations. In addition, due to eigenvalue/vector evaluation on each integration step, its computation times are several times longer than for the other methods (see Table 7.1 in the previous section), known from the theory of ordinary differential equations. This is why the authors of [7.30] need more than 2000 equidistant points of integration for a trapezoidal profile, for which the better adapted numerical integration scheme can suffice with 300 points. Unfortunately, the authors of [7.30] do not consider the differential method as a “reference method” in their work.

Appendix 7.A: S-matrix propagation algorithm

Almost all electromagnetic theories work by providing the link between the electromagnetic field amplitude values established on two different interfaces. These values could be calculated in the real or the inverted space, or the projections of the field on some functional basis, etc. Whatever the theory, if the media are linear, the link can be expressed in a matrix form:

$$A_p = T_p A_{p-1}\,. \tag{7.144}$$

Here, A stands for a column vector containing the field amplitudes in the given basis, the first interface has a number p–1, and the second on, p. T_p is called transmission matrix between the interface (p–1) and p.

Numerical problem arises due to the fact, that the "propagation" between different interfaces contains, in general, both growing and decreasing terms, due to both absorption losses or/and evanescent character of some field components. If a real field term propagates from p–1 to p (the green arrow in Fig.7.A.1), it never grows (unless media with optical gains). Same is valid for the true propagation from interface p to p–1. However, eq. (7.144) is asymmetrical, i.e., it contains propagation only from interface p–1 to p, thus a naturally decreasing field that propagates in the opposite direction (from p to p–1), will be expressed in the T-matrix in the form of growing terms (the red arrow in Fig. 7.A.1). If the propagation length is sufficiently large, these artificial growing terms can overweight the other terms, mainly due to the finite numerical length of the computer word.

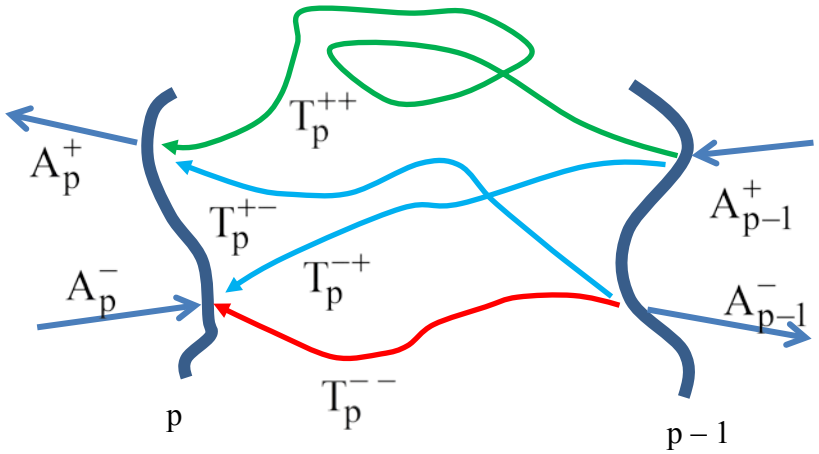


Fig.7.A.1. Schematical representation of the action of the T-matrix between interfaces p-1 and p

One approach that overcomes this problem and that has become quite popular during the last 15 years is the so-called S-matrix propagation algorithm, S staying for 'scattering'. The basic idea is quite simple: As far as the problem of growing terms has been identified, let us try to do as Nature, by determining another link between the field amplitudes, by separating them into terms propagating (or decreasing) in direction (p–1 → p) or in direction (p → p–1). Let us denote the first set with superscript +, and the second set by a superscript –. The S-matrix between the two interfaces provides the following link:

$$\begin{pmatrix} A_p^+ \\ A_{p-1}^- \end{pmatrix} = S_{p,p-1} \begin{pmatrix} A_{p-1}^+ \\ A_p^- \end{pmatrix}. \tag{7.145}$$

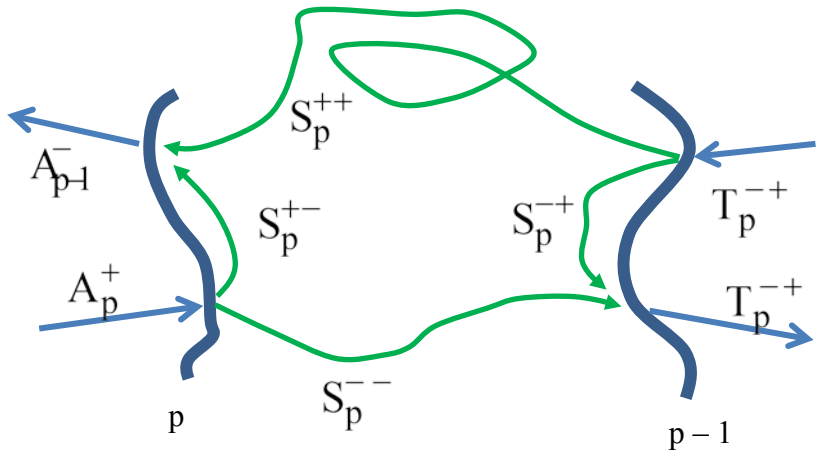


Fig.7.A.2. Action of the S-matrix between interface p-1 and interface p.

The physical meaning is that amplitude A_{p-1}^{+}, which propagates from p–1 to p is defined on p–1 and is not growing in-between p-1 and p. In the same manner, the amplitude A_p^{-} that represents propagation from p to p–1 is defined on the interface p and is not growing in direction of interface p–1. To say in other words, the amplitude A_{p-1}^{+} is incident on the interface p–1 from the previous interface p–2, the second amplitude A_p^{-} is incident on p from p+1, while the amplitudes on the left-hand side of eq.(7.145) are the amplitudes that are scattered in direction to the outside interfaces (p–2 and p+1), thus the name of the scattering matrix S. As observed in Fig.7.A.2., the blocks S_p^{--} and S_p^{++} describe the physically correct transmission between p and p–1 or between p–1 and p, respectively, while the other two blocks, S_p^{-+} and S_p^{-+} describe the reflection on the interface p or p–1, respectively. This interpretation explains why there are no numerical problems due to the growing non-physical interactions when using the S-matrix.

The advantage of this formalism is the absence of artificially growing terms in S. The inconvenience is that electromagnetic theories cannot give a direct expression of the matrix S. However, it is possible to express it by using the T-matrix elements, if it is possible to calculate them correctly. If the 'distance' between interface p–1 and p is quite large (with respect to the growing speed of the growing terms), there is loss of precision in determining the T-matrix. The problem can be solved by introducing additional artificial interfaces between p–1 and p in a such manner that to be able to correctly calculate the T-matrix in each subslice. Once the T-matrix calculated, the S-matrix can be obtained in a closed form. However, the total electromagnetic problem of diffraction (or scattering) requires the knowledge of the entire S-matrix of the system, because the physical problem to be solved needs to express the scattered fields as a function of the fields incident on the system (or generated inside, as is the case for electromagnetic antennas). There exists an iterative algorithm that enables us to establish the total S-matrix without calculating the elementary S-matrix between each consecutive pairs of interfaces, as stated in eq. (7.145). For that sake, we

define another intermediate S-matrix that corresponds to the scattering between some initial interface (numbered as 0) and the interface with number p, $S_p \equiv S_{p,0}$:

$$\begin{pmatrix} A_p^+ \\ A_0^- \end{pmatrix} = S_p \begin{pmatrix} A_0^+ \\ A_p^- \end{pmatrix}. \tag{7.146}$$

The initializing values of S_0 for p = 0 are just the elements of the unity matrix.

As already said, it is necessary to be able to calculate the T-matrices for each intermediate medium between the interfaces. When advancing from the interface p to p+1, we obtain the T-matrix with subscript p+1:

$$\begin{pmatrix} A_{p+1}^+ \\ A_{p+1}^- \end{pmatrix} = T_{p+1} \begin{pmatrix} A_p^+ \\ A_p^- \end{pmatrix}. \tag{7.147}$$

That will be expanded in the form:

$$\begin{pmatrix} A_{p+1}^+ \\ A_{p+1}^- \end{pmatrix} = \begin{pmatrix} T_{p+1}^{++} & T_{p+1}^{+-} \\ T_{p+1}^{-+} & T_{p+1}^{--} \end{pmatrix} \begin{pmatrix} A_p^+ \\ A_p^- \end{pmatrix}. \tag{7.148}$$

It is obvious from the previous considerations that the growing terms are potentially present in the block T_{p+1}^{--} ('antipropagation' from p to p+1), while the blocks T_{p+1}^{++} and T_{p+1}^{-+} can contain decreasing terms ('propagation from p to p+1), i.e., it could be numerically instable to invert them.

On the other hand, the 'next' S-matrix will link the amplitudes with index 0 to the amplitudes (p+1):

$$\begin{pmatrix} A_{p+1}^+ \\ A_0^- \end{pmatrix} = S_{p+1} \begin{pmatrix} A_0^+ \\ A_{p+1}^- \end{pmatrix}. \tag{7.149}$$

Eqs. (7.146)-(7.149) enable us to express the matrix S_{p+1} as a function of S_p and T_{p+1}. At first, we express A_{p+1}^- from eq.(7.148) and substitute A_p^+ from eq.(7.146):

$$A_{p+1}^- = T_{p+1}^{-+} A_p^+ + T_{p+1}^{--} A_p^- = T_{p+1}^{-+} S_p^{++} A_0^+ + \left(T_{p+1}^{-+} S_p^{+-} + T_{p+1}^{--} \right) A_p^-. \tag{7.150}$$

Let us denote as $\mathbb{Z}_{p+1} = \left(T_{p+1}^{-+} S_p^{+-} + T_{p+1}^{--} \right)^{-1}$ in order to eliminate A_p^-:

$$A_p^- = \mathbb{Z}_{p+1} A_{p+1}^- - \mathbb{Z}_{p+1} T_{p+1}^{-+} S_p^{++} A_0^+. \tag{7.151}$$

The next step is to expand the first line of eq.(7.148):

$$\begin{aligned} A_{p+1}^{+} &= T_{p+1}^{++}A_p^{+} + T_{p+1}^{+-}A_p^{-} = T_{p+1}^{++}S_p^{++}A_0^{+} + \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)A_p^{-} \\ &= T_{p+1}^{++}S_p^{++}A_0^{+} + \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)\left(\mathbb{Z}_{p+1}A_{p+1}^{-} - \mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++}A_0^{+}\right) \\ &= \left[T_{p+1}^{++}S_p^{++} - \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)\mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++}\right]A_0^{+} + \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)\mathbb{Z}_{p+1}A_{p+1}^{-} \end{aligned} . \quad (7.152)$$

The comparison with eq.(7.149) gives the first two block-elements of S_{p+1}:

$$S_{p+1}^{+-} = \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)\mathbb{Z}_{p+1} . \quad (7.153)$$

$$\begin{aligned} S_{p+1}^{++} &= T_{p+1}^{++}S_p^{++} - \left(T_{p+1}^{+-} + T_{p+1}^{++}S_p^{+-}\right)\mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++} \\ &= \left(T_{p+1}^{++} - S_{p+1}^{+-}T_{p+1}^{-+}\right)S_p^{++} \end{aligned} . \quad (7.154)$$

From eq.(7.146)

$$\begin{aligned} A_0^{-} &= S_p^{-+}A_0^{+} + S_p^{--}A_p^{-} = S_p^{-+}A_0^{+} + S_p^{--}\left(\mathbb{Z}_{p+1}A_{p+1}^{-} - \mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++}A_0^{+}\right) \\ &= \left(S_p^{-+} - S_p^{--}\mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++}\right)A_0^{+} + S_p^{--}\mathbb{Z}_{p+1}A_{p+1}^{-} \end{aligned} . \quad (7.155)$$

so that

$$S_{p+1}^{--} = S_p^{--}\mathbb{Z}_{p+1} . \quad (7.156)$$

$$\begin{aligned} S_{p+1}^{-+} &= S_p^{-+} - S_p^{--}\mathbb{Z}_{p+1}T_{p+1}^{-+}S_p^{++} \\ &= S_p^{-+} - S_{p+1}^{--}T_{p+1}^{-+}S_p^{++} \end{aligned} . \quad (7.157)$$

These relations exist in several possible forms, but this one is quite well adapted to the case without incident waves on interface 0, because in the iterative algorithm we need to calculate only the half of the blocks, namely the two given by eqs. (7.153) and (7.156).

The only matrix inversion in the iterative algorithm concerns the procedure to obtain the matrix $\mathbb{Z}$. The initial matrix $\mathbb{Z}^{-1}$ contains the potentially large terms from T_{p+1}^{--}, so that its inversion creates neither numerical problems to be inverted, nor growing terms to create numerical instabilities.

Appendix 7.B: Inverted S-matrix propagation algorithm

In Appendix A we have seen how to avoid numerical instabilities due to the artificially growing terms that appear when the propagation of the field amplitudes from one interface to another is made in the wrong direction, a typical property of a half of the field amplitudes used in the transmission matrix approach.

In some cases (for example, the Integral method applied to multilayer grating, but also coordinate transformation method used for a stack containing different profiles), the numerical solution that has been obtained provides a link, having a form inverse to eq.(7.147)

$$\tilde{T}_{p+1}\begin{pmatrix} A_{p+1}^{+} \\ A_{p+1}^{-} \end{pmatrix} = \begin{pmatrix} A_{p}^{+} \\ A_{p}^{-} \end{pmatrix}. \tag{7.158}$$

Of course, it is easy to obtain the form of eq. (7.147) by simply inverting $\tilde{T}_{p+1}$, but better to avoid this, because some blocks of the matrix contain large terms compared to the others. In particular, the block $\tilde{T}_{p+1}^{++}$ is responsible for a physical 'antipropagation' from p+1 to p, so that potentially it contains growing terms (as it was that case with T_{p+1}^{--}) in Appendix A.

We can avoid the direct inversion of $\tilde{T}_{p+1}$ by applying a similar procedure as in Appendix 7.A in order to obtain the S-matrix of the stack. Equation (7.158) is expanded in the form:

$$A_{p}^{+} = \tilde{T}_{p+1}^{++}A_{p+1}^{+} + \tilde{T}_{p+1}^{+-}A_{p+1}^{-}. \tag{7.159}$$

$$A_{p}^{-} = \tilde{T}_{p+1}^{-+}A_{p+1}^{+} + \tilde{T}_{p+1}^{--}A_{p+1}^{-}. \tag{7.160}$$

On the other hand, from eq.(7.146) we have:

$$A_{p}^{+} = S_{p}^{++}A_{0}^{+} + S_{p}^{+-}A_{p}^{-}. \tag{7.161}$$

so that

$$S_{p}^{++}A_{0}^{+} + S_{p}^{+-}A_{p}^{-} = \tilde{T}_{p+1}^{++}A_{p+1}^{+} + \tilde{T}_{p+1}^{+-}A_{p+1}^{-}. \tag{7.162}$$

$$S_{p}^{++}A_{0}^{+} + S_{p}^{+-}\left(\tilde{T}_{p+1}^{-+}A_{p+1}^{+} + \tilde{T}_{p+1}^{--}A_{p+1}^{-}\right) = \tilde{T}_{p+1}^{++}A_{p+1}^{+} + \tilde{T}_{p+1}^{+-}A_{p+1}^{-}. \tag{7.163}$$

and

$$S_{p}^{++}A_{0}^{+} = \left(\tilde{T}_{p+1}^{++} - S_{p}^{+-}\tilde{T}_{p+1}^{-+}\right)A_{p+1}^{+} + \left(\tilde{T}_{p+1}^{+-} - S_{p}^{+-}\tilde{T}_{p+1}^{--}\right)A_{p+1}^{-}. \tag{7.164}$$

Now we can identify half of the blocks of S_{p+1} from eq.(7.149):

$$S_{p+1}^{++} = \tilde{\mathbb{Z}}_{p+1}S_{p}^{++}. \tag{7.165}$$

$$S_{p+1}^{+-} = -\tilde{\mathbb{Z}}_{p+1}\left(\tilde{T}_{p+1}^{+-} - S_{p}^{+-}\tilde{T}_{p+1}^{--}\right). \tag{7.166}$$

with $\tilde{\mathbb{Z}}_{p+1} = \left(\tilde{T}_{p+1}^{++} - S_{p}^{+-}\tilde{T}_{p+1}^{-+}\right)^{-1}$ that contains the numerically dangerous growing terms in $\tilde{T}_{p+1}^{++}$, in the same manner that the matrix $\mathbb{Z}_{p+1}$ in Appendix 7.A 'envelopes' the growing terms in T_{p+1}^{--}.

The other two blocks can be obtained by staring with the identity:

$$A_0^- = S_p^{-+} A_0^+ + S_p^{--} A_p^-, \tag{7.167}$$

and using eq.(7.160) :

$$A_0^- = S_p^{-+} A_0^+ + S_p^{--} \left(\tilde{T}_{p+1}^{-+} A_{p+1}^+ + \tilde{T}_{p+1}^{--} A_{p+1}^- \right). \tag{7.168}$$

When taking into account that two blocks of S_{p+1} are already known and given in eqs. (7.165) and (7.166) , we can eliminate $A_{p+1}^+ = S_{p+1}^{++} A_0^+ + S_{p+1}^{+-} A_{p+1}^-$:

$$A_0^- = \left(S_p^{-+} + S_p^{--} \tilde{T}_{p+1}^{-+} S_{p+1}^{++} \right) A_0^+ + S_p^{--} \left(\tilde{T}_{p+1}^{--} + \tilde{T}_{p+1}^{-+} S_{p+1}^{+-} \right) A_{p+1}^- . \tag{7.169}$$

Thus

$$S_{p+1}^{-+} = S_p^{-+} + S_p^{--} \tilde{T}_{p+1}^{-+} S_{p+1}^{++} . \tag{7.170}$$

$$S_{p+1}^{--} = S_p^{--} \left(\tilde{T}_{p+1}^{--} + \tilde{T}_{p+1}^{-+} S_{p+1}^{+-} \right). \tag{7.171}$$

The expressions are quite similar in form to those obtained in Appendix A. Moreover, they allow avoiding the inversion of $\tilde{T}_{p+1}$.

Finally, there exist a combination of expressions including partial T-matrices, treated separately in Appendix 7.A and 7.B. In some cases the link between the amplitudes on two consecutive interfaces or across a single interface that separates two different media can be expressed in the form:

$$\tilde{\mathfrak{T}}_{p+1} \begin{pmatrix} A_{p+1}^+ \\ A_{p+1}^- \end{pmatrix} = \mathfrak{T}_{p+1} \begin{pmatrix} A_p^+ \\ A_p^- \end{pmatrix}. \tag{7.172}$$

Such is the case of the Fourier-modal (RCW) method across each interface, with the partial transmission matrices $\tilde{\mathfrak{T}}_{p+1}$ containing the eigenvectors of the proper modes inside each media. The same expression is obtained in the coordinate transformation method when using eigenvalue technique of integration. Usually, in both approaches, one obtains the full transmission matrix by inverting $\tilde{\mathfrak{T}}_{p+1}$ and multiplying the result by $\mathfrak{T}_{p+1}$. If this creates numerical problems (for thick layers), such direct approach is not applicable. In that case it is better advised to apply twice the S-matrix algorithm, at first in each direct form (Appendix 7.A), and then in the currently discussed inverted form. It is quite easy to understand the logic, by introducing a virtual set of amplitudes in eq.(7.172):

$$\begin{pmatrix} \tilde{A}_{p+1}^+ \\ \tilde{A}_{p+1}^- \end{pmatrix} = \mathfrak{T}_{p+1} \begin{pmatrix} A_p^+ \\ A_p^- \end{pmatrix}, \tag{7.173}$$

$$\tilde{\mathfrak{T}}_{p+1} \begin{pmatrix} A_{p+1}^+ \\ A_{p+1}^- \end{pmatrix} = \begin{pmatrix} \tilde{A}_p^+ \\ \tilde{A}_p^- \end{pmatrix}. \tag{7.174}$$

References:

7.1.a. N. Bonod, E. Popov, M. Neviere: "Light transmission through a subwavelength microstructured aperture: electromagnetic theory and applications," Opt. Commun. 245, 355-361 (2005)

7.1.b. P. Boyer, E. Popov, M. Nevière, and G. Renversez: "Diffraction theory: application of the fast Fourier factorization to cylindrical devices with arbitrary cross section lighted in conical mounting," J. Opt. Soc. Am. A 23, 1146-1158 (2006)

7.1.c. S. Campbell, R. C. McPhedran, C. M. de Sterke, and L. C. Botten, "Differential multipole method for microstructured optical fibers," J. Opt. Soc. Am. B **21**, 1919-1928 (2004)

7.1.d P. Boyer, E. Popov, G. Renversez, and M. Nevière, "A new differential method applied to the study of arbitrary cross section microstructured optical fibers," Opt. Quant. Electron. **38**, 217-230 (2006)

7.2. B. Stout, M. Nevière, and E. Popov: "Mie scattering by an anisotropic object. Part II: Arbitrary-shaped object – differential theory," J. Opt. Soc. Am. A **23**, 1124-1134 (2006)

7.3. M. A. Melkanoff, T. Sawada, and J. Raynal, "Nuclear optical model calculations," in *Methods in Computationnal Physics*, **1**, 1-80, (Academic Press, New York, 1966)

7.4. G. Cerutti-Maori, R. Petit, and M. Cadilhac, "Etude Numérique du champ diffracté par un réseau," C. R. Ac. Sc. Paris **268**, 1060-1063 (1969)

7.5. M. Nevière, M. Cadilhac, and R. Petit, "Applications of conformal mapping to the diffraction of electromagnetic waves by grating," IEEE Trans. Ant. Propag. AP-21, 37-46 (1973)

7.6.a. M. Nevière, R. Petit, and M. Cadilhac, "About the theory of optical grating coupler-waveguide systems," Opt. Commun. **8**, 113-117 (1973)

7.6b. M. Nevière, P. Vincent, R. Petit, and M. Cadilhac, "Systematic study of resonances of holographic thin film couplers," Opt. Commun. **9**, 48-53 (1973)

7.7.a. M. Nevière, G. Cerutti-Maori, and M. Cadilhac, "Sur une nouvelle méthode de résolution du problème de la diffraction d'une onde plane par un réseau infiniment conducteur," Opt. Commun. 3, 48-52 (1971)

7.7.b. M. Nevière, P. Vincent, and R. Petit, "Sur la théorie du réseau conducteur et ses applications à l'optique," Nouv. Rev. Opt. 5, 65-77 (1974)

7.8. P. Vincent, "Differential methods," in *Electromagnetic Theory of Gratings*, R. Petit, ed. (Springer-Verlag Berlin, 1980), ch. 4

7.9. G. Tayeb, Thèse "Contribution à l'étude de la diffraction des ondes électromagnétiques par des réseaux. Reflexion sur les méthodes existantes et sur leur extension aux milieux anisotropes," Université Aix-Marseille III (1990)

7.10.a. F. Montiel and M. Nevière, "Differential theory of gratings: extention to deep gratings of arbitrary profile and permittivity through the R-matrix propagation algorithm," J. Opt. Soc. Am. A **11**, 3241-3250 (1994)

7.10.b. N. Chateau and J. P. Hugonin, "Algorithm for the rigorous coupled-wave analysis of grating diffraction," J. Opt. Soc. Am. A **11**, 1321-1331 (1994)

7.10.c. L. Li, "Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings," J. Opt. Soc. Am. A **13**, 1024-1035 (1996)

7.11.a. P. Lalanne and G. M. Morris, "Highly improved convergence of the coupled-wave method for TM polarization," J. Opt. Soc. Am. A **13**, 779-784 (1996)

7.11.b. G. Granet and B. Guizal, "Efficient implementation of the coupled-wave method for metallic gratings in TM polarization," J. Opt. Soc. Am. A **13**, 1019-1023 (1996)

7.12. L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," J. Opt. Soc. Am. A **13**, 1870-1876 (1996)

7.13. E. Popov and M. Nevière, "Grating theory: new equations in Fourier space leading to fast converging results for TM polarization," J. Opt. Soc. Am. A **17**, 1773-1784 (2000)

7.14. M. Nevière and E. Popov, *Light Propagation in Periodic Media: Differential Theory and Design* (Marcel Dekker, New York, Basel, 2003)

7.15. M. Cadilhac, "Some mathematical aspects of the grating theory," in *Electromagnetic Theory of Gratings*, R. Petit ed. (Springer-Verlag Berlin, 1980)

7.16. C. H. Wilcox, "Scattering theory for the D'Alembert equation in exterior domains," in *Lecture Notes in Mathemetics*, vol. 442, (Springer, Berlin, 1975)

7.17. L. Li, "Fourier modal method for crossed anisotropic gratings with arbitrary permittivity and permeability tensors," J. Opt. A: Pure Appl. Opt. **5**, 345-355 (2003)

7.18. T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff, "Extraordinary optical transmission through subwavelength hole arrays," Nature **391**, 667-669 (1998)

7.19. L. Li, "Oblique-coordinate-system-based Chandezon method for modeling one-dimensinally periodic, multilayer, inhomogeneous, anisotropic gratings," J. Opt. Soc. A **16**, 2521-2531 (1999)

7.20. Th. Schuster, J. Ruoff, N. Kerwien, S. Rafler, and W. Osten, "Normal vector method for convergence improvement using the RCWA for crossed gratings," J. Opt. Soc. Am. A **24**, 2880-2890 (2007)

7.21. P. Götz, Th. Schuster, K. Frenner, S. Rafler, and W. Osten, "Normal vector method for the RCWA with automated vector field generation," Opt. Express **16**, 17295-17301 (2008)

7.22. J. Bischoff, "Formulation of the normal vector RCWA for symmetric crossed gratings in symmetric mountings," J. Opt. Soc. A **27**, 1024-1031 (2010)

7.23. L. Li, "New formulation of the Fourier modal method for crossed surface-relief gratings," J. Opt. Soc. Am. A **14**, 2758-2767 (1997)

7.24. Th. Weiss, G. Granet; N. Gippius, S. Tikhodeev, and H. Giessen, "Matched coordinates and adaptive spatial resolution in the Fourier modal method," Opt. Express **17**, 8051-8061 (2009)

7.25. W. Press, S. Teulkolsky, W. Vetterling, and B. Flannery: Numerical Recipes, The art of Scientific Computing, Third Edition (Cmbridge Univ. Press 2007), see ch.17.5.

7.26. J. Butcher, *Numerical Methods for Ordinary Differential Equations* (John Wiley, 2003)

7.27. A. Quarteroni, R. Sacco, F. Saleri, *Matematica Numerica*, (Springer Verlag,2000)

7.28 A. Iserles, ed.: *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 1996

7.29. E. Popov, M. Nevière, B. Gralak, and G. Tayeb, "Staircase approximation validity for arbitrary-shaped gratings," J. Opt. Soc. Am. A **19**, 33-42 (2002)

7.30. I. Gushchin and A. Tishchenko, "Fourier modal method for relief gratings with oblique boundary conditions," J. Opt. Soc. Am. A **27**, 1575-1583 (2010)

7.31. L. Li, "Fourier Modal Method," ch.13 of *Gratings: Theory and Numeric Applications, Second Edition Revisited*, Ed. E. Popov (AMU, Marseille, 2014)

Chapter 8:

Coordinate Transformation Methods

Gérard Granet

Table of Contents:

Chapter 8

Coordinate Transformation Methods

Gérard Granet

Clermont Université, Université Blaise Pascal, Institut Pascal, F-63000 Clermont-Ferrand, France
CNRS, UMR 6602, Institut Pascal, F-63177 Aubière, France
Gerard.GRANET@lasmea.univ-bpclermont.fr

8.1 Introduction

The C-method was born in the eighties in Clermont-Ferrand , France, from the need to solve rigorously diffraction problems at corrugated periodic surfaces in the resonance regime [1], [2], [3]. The main difficulty of such problems is the matching of boundaries conditions. It is obvious that any method aimed at solving Maxwell's equation is all the more efficient since it is able to fit the geometry of the problem. For that purpose, Chandezon et al introduced the so called translation coordinate system deduced from the Cartesian coordinate system $x,y,z)$ by the relations $x = x^1$, $y = x^2$, $z = x^3 + a(x^1)$ where $a(x^1)$ is a continuously differentiable function describing the surface profile. Hence since the boundary of the physical problem coincides with coordinate surfaces, writing boundary conditions is as simple as it is for classical problems in Cartesian, cylindrical, or spherical coordinates . This is the first ingredient of C-method. The second one is to write Maxwell's equation under the covariant form. This formulation comes from relativity where the use of curvilinear non orthogonal coordinate system is essential and natural. The main feature of this formalism is that Maxwell's equations remain invariant in any coordinate system, the geometry being shifted into the constitutive relations. Chandezon et al derived their 3D formulation from the general 4D relativistic Post's formalism [4] and evidently used tensorial calculus. Although it is with no doubt the most elegant and efficient way to deal with electromagnetic in general curvilinear coordinates it is also probably the reason why the theory appeared difficult to understand to many scientists. The third ingredient of C–method is that it is a modal method. This nice property is linked with the translation coordinate system in which a diffraction problem may be expressed as an eigenvalue eigenvector problem with periodic boundary conditions. The last feature of C-method is the numerical method of solution. The matrix operator is obtained by expanding field components into Floquet-Fourier harmonics and by projecting Maxwell's equations onto periodic exponential functions. The above four features may be resumed by saying that C-method is a curvilinear coordinate modal method by Fourier expansion [5]. Since the original papers, The C-method has gone through many stages of extension and improvement. The original theory was formulated for uncoated perfectly conducting gratings in classical mount. Various authors extended the method to conical diffraction mountings [6],[7]. Granet et al [11], Li et al [12] and Preist et al [13] allowed the various profiles

of a stack of gratings to be different from each other, although keeping the periodicity. Solving the vertical faces case in a simple manner, Plumey et al [14] have shown that the method can be applied to overhanging gratings Preist et al obtained the same results by applying the usual coordinate transformation to oblique coordinates [15]. In the numerical context, Li [16] and Cotter et al [17] improved the numerical stability of the C-method by using the S-matrix propagation algorithm for multilayer gratings. It is seen that C-method has been applied to a large class of surface relief gratings and multilayer coated gratings. The key point of C-method is the joint use of curvilinear coordinates and covariant formulation of Maxwell's equations. All the new developments in the modelling of gratings like Adaptive Spatial Resolution [18],[19],[20], and Matched coordinate [21]derive from this fundamental observation.

8.2 C-Method

In Euclidean space with origin O and basis vector $e_x, e_y, , e_z, ,$ let us consider an infinite cylindrical surface (Σ) whose elements are parallel to the y axis. This surface separates two linear homogeneous and isotropic media denoted (1) and (2). In Cartesian coordinates such a surface can be described by equation $z = a(x)$. Any electromagnetic field interacting with this particular geometry satisfies some boundary conditions. For instance, the tangential components of the electric field vector and the normal component of the displacement field vector are continuous at the surface. The point is that boundary conditions involve quantities that obviously depend on the position at which they are considered on the surface. We are thus led to look for a coordinate system which fits the problem and makes it more readily solvable than it is in a Cartesian framework. The so-called translation coordinate system (x^1, x^2, x^3) introduced by Chandezon and defined from the Cartesian coordinate system by the direct transformation (curvilinear coordinates to Cartesian coordinates) :

$$x = x^1, \qquad y = x^2, \qquad z = x^3 + a(x^1) \tag{8.1}$$

or the inverse transformation (Cartesian coordinates to curvilinear coordinates):

$$x^1 = x, \qquad x^2 = y, \qquad x^3 = z - a(x) \tag{8.2}$$

is one such system. It makes the surface (Σ) coincide with the coordinate surface $x^3 = 0$. A point $M(x, y, z = a(x)$ at the surface (Σ) is now referenced by the triplet $(x^1, x^2, 0)$. The coordinate surface $x^3 = x_0^3$ is obtained by translating each point at surface (Σ) with vector $x_0^3 e_z$, hence the name given by Chandezon to this particular coordinate system: translation coordinate system. The change of coordinates may also be considered as a change of variable. This view

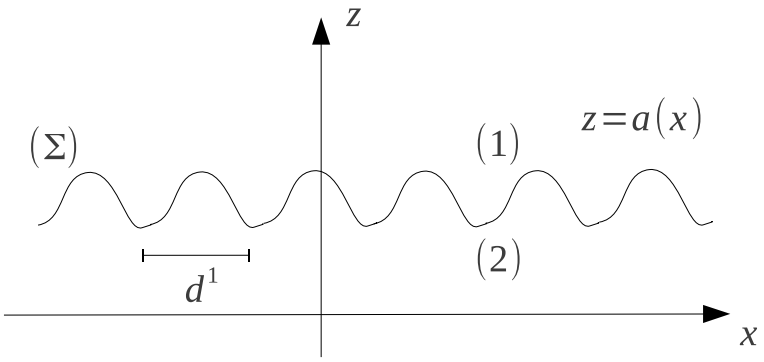


Figure 8.1: Geometry of the problem: Two media are separated by a cylindrical periodic surface, with period d^1 described by the equation z=a(x)

.

point allows a better understanding of the numerical behaviour of C-method and its connection with Rayleigh expansions. There is actually no difference in the way of deriving the elementary solutions of the scalar Helmholtz equation in Cartesian coordinates or in translation coordinate systems. Both are eigenvectors of an eigenvalue problem with pseudo-periodic boundary conditions. In both cases, the operator eigenvalue problem is transformed into a matrix eigenvalue problem thanks to the Galerkin method with pseudo periodic functions as expansion and test functions. Hence solving the scalar Helmholtz equation in any coordinate system is the very first step when implementing C-method. In the next paragraphs we shall focus on this issue before solving a grating problem.

8.2.1 Modal equation in the Cartesian coordinate system

Consider an homogeneous region with relative, possibly complex, permittivity, ε. In the harmonic regime with a time dependence of $\exp(-i\omega t)$, it is possible to construct general solutions to the field equations once we have general solutions to the scalar Helmholtz equation. So as a first task, we are going to investigate elementary solutions to the Helmholtz equation written in the translation coordinate system. Let us start from Cartesian coordinates in which 2D scalar Helmholtz equation is

$$\left(\partial_x^2 + \partial_z^2 + k^2\right) \mathscr{F} = 0 \tag{8.3}$$

where $k = \omega\sqrt{\mu_0\varepsilon}$ is the wave-number. The coefficients of the Helmholtz equation are independent of z so we seek solutions of the form $\mathscr{F}(x,z) = \exp(i\gamma z)F(x)$ The Helmholtz equation becomes:

$$(\partial_x^2 + k^2)F(x) = \gamma^2 F(x) \tag{8.4}$$

Function $F(x)$ is thus an eigenmode of equation (8.4). The requirement that the eigenmodes satisfy the pseudo-periodicity condition $F(x+d^1) = \exp(i\alpha_O d^1)F(x)$ is automatically fulfilled by their expansion into Floquet-Fourier series:

$$F(x) = \sum_{-\infty}^{+\infty} F_m \exp(i\alpha_m x) \tag{8.5}$$

$\alpha_m = \alpha_0 + mK_1$, $K_1 = \dfrac{2\pi}{d^1}$, $m \in \mathbb{N}$ and α_0 is some real parameter. By introducing (8.5) into (8.3) and by projecting onto pseudo-periodic functions $\exp(i\alpha_n x)$, one obtains the matrix equation:

$$\gamma^2 \boldsymbol{F} = \left[k^2 \boldsymbol{I} - \boldsymbol{\alpha}\right] \boldsymbol{F} \tag{8.6}$$

where $\boldsymbol{F}$ is a column vector whose elements are the F_m and $\boldsymbol{\alpha}$ is a diagonal matrix w hose elements are the α_m and $\boldsymbol{I}$ is the identity matrix. The solution to the above matrix eigenvalue equation is of course trivial since the matrix is diagonal. Let us introduce subscript q to number the eigenvalues and the eigenfunctions. The eigenvalues γ_q are deduced from their squared number:

$$\gamma_q^2 = k^2 - \alpha_q^2 \tag{8.7}$$

and the eigenvectors are determined by $F_{mq} = \delta_{mq}$ where δ_{mq} is the Kronecker symbol. The square root of γ_q^2 is defined as follows:

$$\gamma_q = \begin{cases} \sqrt{\gamma_q^2} & \text{if} \quad \gamma_q^2 \in \mathbb{R}^+ \\ \sqrt{-\gamma_q^2} & \text{if} \quad \gamma_q^2 \in \mathbb{R}^- \\ \left(\gamma_q^2\right)^{1/2} & \text{with positive imaginary part if} \quad \gamma_q^2 \in \mathbb{C} \end{cases} \tag{8.8}$$

Finally , $\mathscr{F}(x,z)$ can be represented by superposition of eigenmodes

$$\mathscr{F}(x,z) = \mathscr{F}^+(x,z) + \mathscr{F}^-(x,z) \tag{8.9}$$

with

$$\mathscr{F}^{+}(x,z) = \sum_{q=-\infty}^{q=+\infty} A_q^{+} \exp(i\gamma_q z) \sum_{m=-\infty}^{m=+\infty} \delta_{mq} \exp(i\alpha_m x) \tag{8.10}$$

$$\mathscr{F}^{-}(x,z) = \sum_{q=-\infty}^{q=+\infty} A_q^{-} \exp(-i\gamma_q z) \sum_{m=-\infty}^{m=+\infty} \delta_{mq} \exp(i\alpha_m x) \tag{8.11}$$

There are two sets of modes, the number of which are equal: those propagating or decaying in the positive direction of z and those propagating or decaying in the opposite direction. We denote these modes by superscript $+$ and $-$respectively. The z dependence of an eigenmode is determined by function $\exp(i\gamma_p z)$ By increasing z to $z+\Delta z$, $\exp(i\gamma_q z)$ is multiplied by $\exp(i\gamma_q \Delta z) = \exp(i\Re(\gamma_q)\Delta z) \times \exp(-\Im(\gamma_q)\Delta z)$ The real eigenvalues have $\Im(\gamma_q) = 0$ and correspond therefore to forward modes if $\Re(\gamma_q) > 0$ or backward modes if $\Re(\gamma_q) < 0$. The complex eigenvalues modes have a non-zero imaginary part and possibly also a non-zero real part. The associated eigenmodes decay forward if $\Im(\gamma_q) > 0$ or backward if $\Im(\gamma_q) > 0$. These expansions are known as Rayleigh expansions; they are linear combination of eigenvectors that we call hereafter Rayleigh eigenvectors R_q:

$$R_q(x) = \sum_{m=-\infty}^{m=+\infty} \delta_{mq} \exp(i\alpha_m x) \tag{8.12}$$

In Cartesian coordinates, the solutions to the Helmholtz equations may be regarded as the eigenvectors of a matrix equation. The eigenvalues γ_m are determined by the periodic lateral boundary conditions of the problem and are obtained analytically since the matrix is diagonal. The translation coordinate system preserves the z translation symmetry and also periodic lateral boundary conditions. We may then expect a great formal similitude between solutions obtained in each coordinate system.

8.2.2 Modal equation in terms of the new variables

In this section, we derive the master equation of C-method by considering the change of coordinates as a change of variables. For the change of variables $x^1 = x$, $x^2 = y$, $x^3 = z - a(x)$ the chain rule for derivatives has the form:

$$\begin{cases} \partial_x = \partial_1 - \dot{a}\partial_3 \\ \partial_y = \partial_2 \\ \partial_z = \partial_3 \end{cases} \tag{8.13}$$

Substituting the derivatives (8.13) into (8.3) gives:

$$\left((1+\dot{a}\dot{a})\partial_3^2 - \dot{a}\partial_1\partial_3 - \partial_1\dot{a}\partial_3 + \partial_1^2 + k^2\right)\mathscr{F}(x^1,x^3) = 0 \tag{8.14}$$

the solution of which are the same as the solutions of (8.13) expressed in terms of the new variables.

$$\begin{aligned} \mathscr{F}_a^{+}(x^1,x^3) &= \mathscr{F}^{+}(x = x^1, x^3 = z - a(x)) \\ &= \sum_{q=-\infty}^{q=+\infty} A_q^{+} \exp(i\gamma_q x^3) \sum_{m=-\infty}^{m=+\infty} \delta_{mq} \exp(i\gamma_q a(x^1)) \exp(i\alpha_m x^1) \end{aligned} \tag{8.15}$$

$$\begin{aligned}\mathscr{F}_a^-(x^1,x^3) &= \mathscr{F}^-(x=x^2,x^3=z-a(x)) \\ &= \sum_{q=-\infty}^{q=+\infty} A_q^- \exp(-i\gamma_q x^3) \sum_{m=-\infty}^{m=+\infty} \delta_{mq} \exp(-i\gamma_q a(x^1)) \exp(i\alpha_m x^1)\end{aligned} \tag{8.16}$$

The subscript a indicates the profile dependence of function $\mathscr{F}$.

We call function $\exp\left(\pm i\gamma_q a(x^1)\right)\exp\left(i\alpha_q x^1\right)$ the generalized Rayleigh eigenvector of order q. It is nothing more than plane wave $\exp(\pm i\gamma_q z)exp(i\alpha_q x^1)$ expressed in terms of the new variables x^1 and x^3 and is closely linked with function $a(x^1)$. Let us denote it $R_{a,q}^{\pm}$:

$$R_{a,q}^{\pm} = \exp\left(\pm i\gamma_q a(x^1)\right)\exp(i\alpha_q x^1) = \sum_{m=-\infty}^{m=+\infty} R_{amq}\exp(i\alpha_m x^1) \tag{8.17}$$

It is assumed so far that $a(x^1)$ is periodic with period d^1 hence:

$$\exp(\pm i\gamma_q a(x^1)) = \sum_{p=-\infty}^{p+\infty} L_p^{\pm}\exp\left(\frac{i2\pi p x^1}{d^1}\right) \tag{8.18}$$

with:

$$L_p^{\pm} = \frac{1}{d^1}\int_0^{d^1} \exp\left(\pm i\gamma_q a(x^1)\right)\exp\left(\frac{-i2\pi p x^1}{d^1}\right)dx^1 \tag{8.19}$$

In physical space, the generalized Rayleigh eigenvectors result from the product of a periodic function with a pseudo-periodic one. Thus, in Fourier space, the spectrum of the qth generalized Rayleigh eigenvector is obtained by translating the spectrum of function $\exp(\pm i\gamma_q z)$ with vector $2\pi q/d^1$, that is:

$$R_{amq}^{\pm} = L_{am-q}^{\pm} \tag{8.20}$$

Finally:

$$\mathscr{F}_a^+(x^1,x^3) = \sum_{q=-\infty}^{q=+\infty} A_q^+\exp(i\gamma_q x^3)\sum_{m=-\infty}^{m=+\infty} L_{am-q}^+\exp(i\alpha_m x^1) \tag{8.21}$$

$$\mathscr{F}_a^-(x^1,x^3) = \sum_{q=-\infty}^{q=+\infty} A_q^-\exp(-i\gamma_q x^3)\sum_{m=-\infty}^{m=+\infty} L_{am-q}^-\exp(i\alpha_m x^1) \tag{8.22}$$

Functions (8.21) and (8.22) give the general solution to (8.14). Indeed each element of this solution is a generalized Rayleigh eigenvector associated to index q such that $\gamma_q^2+\alpha_q^2=k^2$ and thus satisfies(8.14). The reason for that is obvious. It is obtain from (8.12) in which we have introduced the same change of variable as the one that has allowed us to get (8.14) from (8.3). From a practical view point, one can only manipulate finite size expansions and it does not make sense to speak of $R_{a,q}^{\pm}(x^1)$. That is why one may wonder if a generalized Rayleigh eigenvector is still a valid a solution of (8.14) when only a finite number of spatial Fourier harmonics is retained to represent it. Let us assume for a while the answer is yes and examine the involvements of such a claim. Introducing an integer M, hereafter denoted truncation number, and letting m run from $-M$ to M the truncated generalized Rayleigh eigenvector writes:

$$R_{a,q}^{\pm(M)}(x^1) = \sum_{m=-M}^{m=+M} R_{a,mq}^{\pm}\exp(i\alpha_m x^1) \tag{8.23}$$

Substituting ∂_3 with $i\gamma_q$, we have:

$$-(1+\dot{a}\dot{a})\gamma_q \dot{R}_a^{\pm(M)} - i(\dot{a}\partial_1 + \partial_1\dot{a})\gamma_q R_a^{\pm,(M)} + (\partial_1^2 + k^2)R_a^{\pm,(M)} = 0 \tag{8.24}$$

where $\dot{R}_a^{\pm(M)}$ denotes $\gamma_q R_a^{\pm,(M)}$. Replacing $\dot{a}$ by the coefficients of its Fourier series $\dot{a}_p$ and denoting $\dot{\boldsymbol{a}}$ the toeplitz matrix whose elements $\dot{a}_{mp}$ are the $\dot{a}_{m-p}$, it is easy to see that the matrix form of relation 8.24 is:

$$\begin{bmatrix} k^2\boldsymbol{I} - \boldsymbol{\alpha}^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{aq}^{\pm} \\ \dot{\boldsymbol{R}}_{aq}^{\pm} \end{bmatrix} = \gamma_q \begin{bmatrix} -\dot{\boldsymbol{a}}\boldsymbol{\alpha} - \boldsymbol{\alpha}\dot{\boldsymbol{a}} & \boldsymbol{I} + \dot{\boldsymbol{a}}\dot{\boldsymbol{a}} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{aq}^{\pm} \\ \dot{\boldsymbol{R}}_{aq}^{\pm} \end{bmatrix} \tag{8.25}$$

where $\boldsymbol{R}_{aq}^{\pm}$ and $\dot{R}_{aq}^{\pm}$ are column vectors formed by the $2M+1$ Fourier coefficients of $R_{aq}^{\pm(M)}$ $\dot{R}_{aq}^{\pm(M)}$ respectively. (8.25) shows that γ_q and $\begin{bmatrix} \boldsymbol{R}_{aq}^{\pm} \\ \dot{\boldsymbol{R}}_{aq}^{\pm} \end{bmatrix}$ are an eigenvalue and an eigenvector of the generalized matrix eigenequation $\boldsymbol{A}\psi = \rho\boldsymbol{B}\psi$. Since $R_{a,q}^{\pm}$ is an exact eigenvector of (8.14) its truncated part can only approximate the solution of (8.14) and consequently, mathematically speaking γ_q cannot be an eigenvalue of (8.25). It follows that our claim was false. Nevertheless, elementary pseudo-periodic solutions to (8.14) do exist and we will derive them in the next paragraph.

8.2.3 Fourier expansion of elementary waves in the translation coordinate system

In this paragraph, we derive the generalized eigenvalue eigenvector matrix equation starting from (8.14) the only assumption being the pseudo periodicity of the field and we discuss the obtained solutions. First, the propagation equation is rewritten as a pair of first-order equations:

$$\begin{bmatrix} k^2 + \partial_1^2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathscr{F} \\ \partial_3\mathscr{F} \end{bmatrix} = -\partial_3 \begin{bmatrix} \dot{a}\partial_1 + \partial_1\dot{a} & 1 + \dot{a}\dot{a} \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \mathscr{F} \\ \partial_3\mathscr{F} \end{bmatrix} \tag{8.26}$$

The coefficients of this equation do not depend on x^3 which allows to write the x^3 dependence as $\exp(i\rho x^3)$. The parameter ρ depends on the boundary conditions that $F(x^1,x^3)$ has to satisfy along x^1 direction. For gratings, periodic with period d^1 along x^1, $\mathscr{F}(x^1+d^1,x^3) = \exp(i\alpha_0 d^1)\mathscr{F}(x^1,x^3)$ where α_0 is some real parameter. $\partial_3\mathscr{F}$ verifies of course the same property. The above requirements on the solution are all fulfilled by expanding function $\mathscr{F}$ and $\partial_3\mathscr{F}$ under the form:

$$\mathscr{F}(x^1,x^3) = \exp(i\rho x^3)F_a(x^1) = \exp(i\rho x^3)\sum_{m=-M}^{m=M} F_{a,m}\exp(i\alpha_m x^1) \tag{8.27}$$

$$\partial_3\mathscr{F}(x^1,x^3) = \exp(i\rho x^3)\dot{F}_a(x^1) = \exp(i\rho x^3)\sum_{m=-M}^{m=M} \dot{F}_{a,m}\exp(i\alpha_m x^1) \tag{8.28}$$

Introducing the above expansions into (8.26) and projecting the latter onto $\exp\left(\dfrac{i2\pi n x^1}{d^1}\right)$ basis, we get the sought algebraic matrix eigenvalue equation from which eigenvalues ρ_q and eigenvectors $\boldsymbol{F}_{a,q}$ are readily obtained thanks to standard computer libraries:

$$\begin{bmatrix} k^2\boldsymbol{I} - \boldsymbol{\alpha}^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{a,q} \\ \dot{\boldsymbol{F}}_{a,q} \end{bmatrix} = \rho_{a,q} \begin{bmatrix} -\dot{\boldsymbol{a}}\boldsymbol{\alpha} - \boldsymbol{\alpha}\dot{\boldsymbol{a}} & \boldsymbol{I} + \dot{\boldsymbol{a}}\dot{\boldsymbol{a}} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{a,q} \\ \dot{\boldsymbol{F}}_{a,q} \end{bmatrix} \tag{8.29}$$

As in the Cartesian coordinate system, it is observed numerically that there are two sets of modes, the number of which are equal: those propagating or decaying in the positive x^3 direction and those propagating or decaying in the opposite direction. Furthermore, it has been shown numerically and analytically [8], that, as the truncation number increases, the computed real eigenvalues converge to the real Rayleigh eigenvalues $\pm\gamma_q$.

$$\lim_{M\to\infty} \pm\rho_{a,q}^{M} = \pm\gamma_q^R \tag{8.30}$$

In the above relation, we have added an extra subscript M to indicate the truncation dependence. Indeed, the truncation order M has to be chosen large enough so that the computed real eigenvectors coincide with a great accuracy with their Rayleigh counterpart. In that case, provided that the eigenvalues are not degenerated, up to a multiplicative constant coefficient, the associated computed eigenvectors tend to the corresponding plane waves expressed in terms of the new variables (x^1, x^2, x^3).

$$\lim_{M\to\infty} F_{a,q}^{\pm(M)} = R_{a,q}^{\pm} \tag{8.31}$$

Thus in the translation coordinate system defined by $x^1 = x$, $x^2 = y$, $x^3 = z - a(x)$ as in the Cartesian coordinate system $Oxyz$, linear combinations of elementary solutions to the Helmholtz equation allow us to express electromagnetic field while giving it a physical meaning in terms of forward and backward waves.We write numerically the solution to the Helmholtz equation as:

$$\mathscr{F}_a^+(x^1,x^3) = \sum_{q\in U^+} A_q^+ \exp(i\rho_{a,q}^+ x^3) R_{a,q}^{+(M)}(x^1) + \sum_{q\in V^+} A_q^+ \exp(i\rho_{a,q}^+(x^3) F_{a,q}^+(x^1) \tag{8.32}$$

$$\mathscr{F}_a^-(x^1,x^3) = \sum_{q\in U^+} A_q^- \exp(i\rho_{a,q}^-(x^3) R_{a,q}^{-(M)}(x^1) + \sum_{q\in V^+} A_q^- \exp(i\rho_{a,q}^-(x^3) F_{a,q}^-(x^1) \tag{8.33}$$

with:

$$F_{a,q}^{\pm}(x^1) = \sum_{m=-M}^{m=+M} F_{a,mq}^{\pm} \exp(i\alpha_m x^1) \tag{8.34}$$

$U^{\pm}$, $V^{\pm}$ denote the sets of indices for the propagating and decaying orders in the positive and negative direction respectively.

$$U^+ = \left\{q/\Re(\rho_{a,q}) > 0 \text{ and } \Im(\rho_{a,q}) = 0\right\} \qquad U^- = \left\{q/\Re(\rho_{a,q}) < 0 \text{ and } \Im(\rho_{a,q}) = 0\right\} \tag{8.35}$$

$$V^+ = \left\{q/\Im(\rho_{a,q}) > 0\right\} \qquad V^- = \left\{q/\Im(\rho_{a,q}) < 0\right\} \tag{8.36}$$

8.3 Application to a grating problem

Let's come back to the one-dimensional grating problem. Consider the electromagnetic problem in which two homogeneous non magnetic media are separated by a cylindrical periodic surface with period d^1 which is invariant along the y axis in the Cartesian coordinate system $Oxyz$. Such a surface, described by equation $z = a(x)$ is illuminated from above by a unit amplitude linear polarized monochromatic plane wave with vacuum wavelength λ_0, angular frequency ω and vacuum wave number $k_0 = 2\pi/\lambda_0$. The wave vector is inclined at θ to the Oz axis. Medium (1)

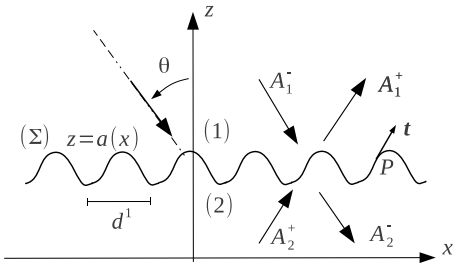


Figure 8.2: Geometry of the diffraction problem. Sketch of the coefficients for scattering matrix

and medium (2) have relative permittivity ε_1 and ε_2 respectively. Time dependence is expressed by the factor $\exp(-i\omega t)$ Such a problem is reduced to the study of the two fundamental cases of polarisation and the unknown function $\mathscr{F}(x,z)$ is the *y* component of the electric or the magnetic field for TE and TM polarization respectively. We solved half the problem since we already determined the general solution to the scalar Helmholtz equation as a linear combination of elementary waves the coefficients of which remain to calculate. The situation is very common in electromagnetic theory: the fields on both side of the grating are expanded in terms of the modes in the respective regions with unknown coefficients. A method of solution known as mode-matching method was developed in the context of guided waves in the micro-wave range. The grating may be considered as a generalized multi-port whose inputs are excited by waves that propagate or decay towards it giving rise to a response at the outputs that consists of the waves that propagate or decay away from it [25],[24]. The mode coupling is caused by the modulation of the interface and by the different constitutive parameters in either side of it. The so-called scattering matrix S_a defined as

$$\begin{bmatrix} \boldsymbol{A}^{(1)+} \\ \boldsymbol{A}^{(2)-} \end{bmatrix} = \boldsymbol{S}_a \begin{bmatrix} \boldsymbol{A}^{(1)-} \\ \boldsymbol{A}^{(2)+} \end{bmatrix} \tag{8.37}$$

provides a linear relation between the output and input coefficients. In a grating problem, the vector formed by the amplitudes of the incoming waves has only one non null component: that corresponding to the incident wave which was assumed enforced to one. The subscript a indicates that the S matrix depends on the profile function $a(x)$. We call $\boldsymbol{S}_a$ matrix an interface scattering matrix. The $\boldsymbol{S}_a$ matrix is derived from boundary conditions at the surface $x^3 = x_0^3$. The change of variable makes it easy to write them. We have solved the scalar Helmholtz equation, the scalar field being a field component tangent to the surface;indeed F coincides with H_y and E_y in TM polarisation and TE polarisation respectively. For simplicity, let us consider TE polarisation where the non null components of the electromagnetic field are E_y, H_x, H_z. Boundary conditions require matching the tangential components of the magnetic and electric field. We have already derived one of them, E_y, we have to derive the tangential component H_t of the electric field given by :

$$H_t = \boldsymbol{H}.\boldsymbol{t} \tag{8.38}$$

where $\boldsymbol{t}$ is the unit vector at point P which is tangential to the grating profile function. It is defined in terms of the $\boldsymbol{e}_x$ and $\boldsymbol{e}_z$ Cartesian unit vectors by:

$$\boldsymbol{t} = \frac{1}{\sqrt{1+\dot{a}^2}}(\boldsymbol{e}_x + \dot{a}\boldsymbol{e}_z) \tag{8.39}$$

The square root in the denominator represents a normalization factor that can be omitted since at a given point, it is identical on both sides of the boundary surface. let us introduce G such

that:

$$\mathscr{G} = iZ\boldsymbol{H}.\boldsymbol{t} \tag{8.40}$$

Where $Z = \sqrt{\mu_0/\varepsilon}$ is the wave impedance. From Maxwells equation we have $i\omega\mu_0 H_x = -\partial_z E_y$ and $i\omega\mu_0 H_z = \partial_x E_y$ thus :

$$\mathscr{G}(x,z) = -\frac{1}{k}(\partial_z \mathscr{F}(x,z) - \dot{a}\partial_x \mathscr{F}(x,z)) \tag{8.41}$$

substituting ∂_3 for ∂_z and $\partial_1 - \dot{a}\partial_3$ for ∂_x we get:

$$\mathscr{G}(x^1,x^3) = -\frac{1}{k}\left((1+\dot{a}\dot{a})\partial_3 - \dot{a}\partial_1\right)\mathscr{F}(x^1,x^3) \tag{8.42}$$

Similarly to $\mathscr{F}$, $\mathscr{G}$ depends on x^3 as $\exp(i\rho x^3)$ and we may write:

$$\mathscr{G}(x^1,x^3) = \exp(i\rho x^3)G(x^1) \tag{8.43}$$

We are now familiar with the operational rules that allow to associate in Fourier space a matrix with an operator. We have:

$$1 + \dot{a}\dot{a} \to \boldsymbol{I} + \dot{\boldsymbol{a}}\dot{\boldsymbol{a}}, \qquad \dot{a}\partial_1 \to i\dot{\boldsymbol{a}}\boldsymbol{\alpha} \tag{8.44}$$

From which we deduce:

$$ik\boldsymbol{G}_a^{\pm} = (I + \dot{\boldsymbol{a}}\dot{\boldsymbol{a}})\,\boldsymbol{F}_a^{\pm}\boldsymbol{\rho}_a - \dot{\boldsymbol{a}}\boldsymbol{\alpha}\boldsymbol{F}_a^{\pm} \tag{8.45}$$

where $\boldsymbol{\rho}$ is a diagonal matrix whose elements are the eigenvalues $\rho_{a,q}$. Writing the continuity of $\mathscr{F}^{(1)}$ and $\mathscr{F}^{(2)}$ and $\mathscr{G}^{(1)}/Z^{(1)}$ and $\mathscr{G}^{(2)}/Z^{(2)}$ at $x^3 = x_0^3$ is straightforward and leads to the following expression of the $\boldsymbol{S}_a$ matrix:

$$\boldsymbol{S}_a = \begin{bmatrix} \boldsymbol{F}_a^{(1)+} & -\boldsymbol{F}_a^{(2)-} \\ \boldsymbol{G}_a^{(1)+} & -\boldsymbol{G}_a^{(2)-} \end{bmatrix}^{-1} \begin{bmatrix} -\boldsymbol{F}_a^{(1)-} & \boldsymbol{F}_a^{(2)+} \\ -\boldsymbol{G}_a^{(1)-} & \boldsymbol{G}_a^{(2)+} \end{bmatrix} \tag{8.46}$$

The knowledge of $\boldsymbol{S}_a$ matrix allows to calculate the constant coefficients of outgoing waves. Since the spectrum of the solutions of the transformed Helmholtz equation include the generalized Rayleigh eigenvectors associated to real Rayleigh eigenvalues the efficiencies may be calculated in the very same way as in the Cartesian coordinate system.

$$R_q = |A_q^{(1)+}|^2 \frac{\gamma_q^{(1)}}{\gamma_0^{(1)}} \qquad T_p = |A_p^{(2)-}|^2 \frac{\gamma_p^{(2)}}{\gamma_0^{(1)}} \tag{8.47}$$

with:

$$\gamma_q^{(1)} = \sqrt{k_0^2\varepsilon_1 - \left(k_0\sqrt{\varepsilon}_1 \sin\theta + q\frac{2\pi}{d^1}\right)^2} \quad \gamma_p^{(2)} = \sqrt{k_0^2\varepsilon_2 - \left(k_0\sqrt{\varepsilon}_1 \sin\theta + p\frac{2\pi}{d^1}\right)^2} \tag{8.48}$$

The values of integers p and q are such that $\gamma_q^{(1)}$ and $\gamma_p^{(2)}$ are real.

8.3.1 Implementation of C-Method

The main interest to consider C-Method through a simple change of variable is to make us understand its numerical link with Rayleigh expansions and to calculate efficiencies as in the Cartesian Coordinate system. Within the framework of translation coordinate systems, starting from Maxwell's equations written under the covariant form, we have shown that all tangential components of the field at surface S could be generated from the longitudinal covariant components along the axis of invariance. Moreover,these components are solutions of the scalar Helmholtz equation. Therefore, one clearly understands that finding the elementary solutions of the scalar Helmholtz equation is the kernel of C-method. To summarize we may enunciate the different steps for solving a grating problem with C-method:

- Define a translation coordinate system.
- Find the elementary waves of the Helmholtz equation. For that purpose use the Galerkin method with $\exp(i\alpha_n x^1)$ as expansion and test functions. Substitute the generalized Rayleigh eigenvectors for the computed eigenvectors associated to real eigenvalues. Sort the elementary waves into forward and backward waves.
- Write boundary conditions at surface Σ and calculate efficiencies as in the Cartesian coordinate system.

8.4 Various formulations of C-method

So far, the Helmholtz equation in the translation coordinate system was derived by using the chain rule for derivatives in the Helmholtz equation written in the translation coordinate system. In this section,we start from the covariant Maxwell's equations and we show that they lead to several operators one of them being the propagation equation. In a homogeneous isotropic medium with permittivity ε and permeability μ, with a time dependence $\exp(-i\omega t)$,the symmetrized Maxwell equations write:

$$\begin{aligned} \xi^{\alpha\beta\gamma}\partial_\beta \mathscr{F}_\gamma &= k\sqrt{g}g^{\alpha\beta}\mathscr{G}_\alpha \\ \xi^{\alpha\beta\gamma}\partial_\beta \mathscr{G}_\gamma &= k\sqrt{g}g^{\alpha\beta}\mathscr{F}_\alpha \end{aligned} \tag{8.49}$$

where $k=\omega\sqrt{\mu\varepsilon}$, the $\mathscr{F}_\gamma$ and the $\mathscr{G}_\beta$ are the complex amplitudes of the electric field and of a renormalized magnetic field respectively. We restrict our analysis to 1D problems in which both the geometry and the solution are independent of y. Practically this means that ∂_2 is null as well as g^{12}, g^{21}, g^{32} and g^{23}. It follows that (8.49) decouple into two fully identical systems where the non null components are $\mathscr{F}_2$, $\mathscr{G}_1$, $\mathscr{G}_3$, and $\mathscr{G}_2$, $\mathscr{F}_1$, $\mathscr{F}_3$ respectively.The first set of three components corresponds to TE polarisation, the second one to TM polarisation. Both polarisations obey the same first order differential equations system written hereafter for TE polarisation:

$$-\partial_3 \mathscr{F}_2 = k\left(\sqrt{g}g^{11}\mathscr{G}_1 + \sqrt{g}g^{13}\mathscr{G}_3\right) \tag{8.50a}$$

$$\partial_1 \mathscr{F}_2 = k\left(\sqrt{g}g^{31}\mathscr{G}_1 + \sqrt{g}g^{33}\mathscr{G}_3\right) \tag{8.50b}$$

$$\partial_3 \mathscr{G}_1 - \partial_1 \mathscr{G}_3 = k\sqrt{g}g^{22}F_2 \tag{8.50c}$$

For TM polarisation, it is enough to permute $\mathscr{F}$ and $\mathscr{G}$. Among the three components of each system, two play a particular role. Let us assume that $x^3 = x_0^3$ separates two isotropic homogeneous media. Then, in TE polarisation $\mathscr{F}_2$ and $\mathscr{G}_1/\sqrt{\frac{\mu}{\varepsilon}}$ have to be continuous at surface $x^3 = x_0^3$. The same conclusions holds for $\mathscr{G}_2/\sqrt{\frac{\mu}{\varepsilon}}$ and $\mathscr{F}_1$ for TM polarisation. So, we have to solve (8.50) for the components labelled by two and by one. C-method is a Fourier based method which means that constitutive relations have to be written in Fourier space. In other words a matrix is to be associated to each element $\sqrt{g}g^{\alpha\beta}$ of the constitutive tensors. The way for doing so should follow the so-called "Fourier factorization" rules derived by Li [22],[23]. let us denote by $(\sqrt{g}g^{\alpha\beta})$ the matrix associated to coefficient $\sqrt{g}g^{\alpha\beta}$. According to Li's rules, the $(\sqrt{g}g^{\alpha\beta})$ write:

$$\begin{aligned}
(\sqrt{g}g^{11}) &= \left[\frac{1}{\sqrt{g}g^{11}}\right]^{-1} \\
(\sqrt{g}g^{13}) &= \left[\frac{1}{\sqrt{g}g^{11}}\right]^{-1}\left[\frac{g^{13}}{g^{11}}\right] \\
(\sqrt{g}g^{31}) &= \left[\frac{g^{31}}{g^{11}}\right]\left[\frac{1}{\sqrt{g}g^{11}}\right]^{-1} \\
(\sqrt{g}g^{33}) &= \left[\frac{1}{\sqrt{g}g^{11}}\right] + \left[\frac{g^{31}}{g^{11}}\right]\left[\frac{1}{\sqrt{g}g^{11}}\right]^{-1}\left[\frac{g^{13}}{g^{11}}\right] \\
(\sqrt{g}g^{22}) &= \left[\sqrt{g}g^{22}\right]
\end{aligned} \tag{8.51}$$

The notation $[f]$ designates the toeplitz matrix whose elements f_{mp} are the f_{m-p} elements of the Fourier series of function $f(x^1)$. For the translation coordinate (x^1, x^2, x^2) such that $x = x^1$, $y = x^2$, $z = x^3 + a(x^1)$, we have:

$$\begin{aligned}
\left(\sqrt{g}g^{11}\right) &\to \boldsymbol{I} \\
\left(\sqrt{g}g^{13}\right) &\to -\dot{\boldsymbol{a}} \\
\left(\sqrt{g}g^{31}\right) &\to -\dot{\boldsymbol{a}} \\
\left(\sqrt{g}g^{33}\right) &\to [\boldsymbol{I} + \dot{\boldsymbol{a}}\dot{\boldsymbol{a}}] \\
\left(\sqrt{g}g^{22}\right) &\to \boldsymbol{I}
\end{aligned} \tag{8.52}$$

In Fourier space, the derivative operator ∂_1 is associated to the diagonal matrix $i\boldsymbol{\alpha}$ the elements of which are the $i\alpha_m$ such that:

$$\alpha_m = \alpha_0 + m\frac{2\pi}{d^1} \tag{8.53}$$

Setting

$$\mathscr{F}_2(x^1, x^3) = \sum_{m=-M}^{m=+M} F_{2m}(x^3)\exp(i\alpha_m x^1) \tag{8.54a}$$

$$\mathscr{G}_1(x^1, x^3) = \sum_{m=-M}^{m=+M} G_{1m}(x^3)\exp(i\alpha_m x^1) \tag{8.54b}$$

$$\mathscr{G}_3(x^1, x^3) = \sum_{m=-M}^{m=+M} G_{3m}(x^3)\exp(i\alpha_m x^1) \tag{8.54c}$$

(8.54d)

We are now able to write (8.50) in Fourier space:

$$-\partial_3 \boldsymbol{F}_2 = k\left(\left(\sqrt{g}g^{11}\right)\boldsymbol{G}_1 + \left(\sqrt{g}g^{13}\right)\boldsymbol{G}_3\right) \tag{8.55a}$$

$$i\boldsymbol{\alpha}\boldsymbol{F}_2 = k\left(\left(\sqrt{g}g^{31}\right)\boldsymbol{G}_1 + \left(\sqrt{g}g^{33}\right)\boldsymbol{G}_3\right) \tag{8.55b}$$

$$\partial_3 \boldsymbol{G}_1 - i\boldsymbol{\alpha}\boldsymbol{G}_3 = k\left(\sqrt{g}g^{22}\right)\boldsymbol{F}_2 \tag{8.55c}$$

where $\boldsymbol{F}_2$, $\boldsymbol{G}_1$ $\boldsymbol{G}_3$ are column vectors of size $2M+1$ whose components are the $F_{2m}(x^3)$, $G_{1m}(x^3)$, $G_{3m}(x^3)$ respectively:

$$\boldsymbol{F}_2(x^3) = [F_{2,-M}(x^3),\, F_{2,-M+1}(x^3),\, \cdots, F_{2,0}(x^3),\, \cdots,\, F_{2,M-1}(x^3),\, F_{2,M}(x^3)]^T \tag{8.56a}$$

$$\boldsymbol{G}_1(x^3) = [G_{1,-M}(x^3),\, G_{1,-M+1}(x^3),\, \cdots,\, G_{1,0}(x^3),\, \cdots,\, G_{1,M-1}(x^3),\, G_{1,M}(x^3)]^T \tag{8.56b}$$

$$\boldsymbol{G}_3(x^3) = [G_{3,-M}(x^3),\, G_{3,-M+1}(x^3),\, \cdots G_{3,0}(x^3),\, \cdots,\, G_{3,M-1}(x^3),\, G_{3,M}(x^3)]^T \tag{8.56c}$$

where the exponent T is for the transposition.

8.4.1 Propagation equation in curvilinear coordinates

From Eqs (8.50a) and (8.50b), $\boldsymbol{G}_1$ and $\boldsymbol{G}_3$ may be expressed in terms of $\boldsymbol{F}_2$

$$k\boldsymbol{G}_1 = -\left(\sqrt{g}g^{33}\right)\partial_3 \boldsymbol{F}_2 + \left(\sqrt{g}g^{13}\right) i\boldsymbol{\alpha}\boldsymbol{F}_2 \tag{8.57}$$

$$k\boldsymbol{G}_3(x^3) = \left(\sqrt{g}g^{31}\right)\partial_3 \boldsymbol{F}_2(x^3) + \left(\sqrt{g}g^{11}\right) i\boldsymbol{\alpha}\boldsymbol{F}_2(x^3) \tag{8.58}$$

(8.50c), in which we substitute $\boldsymbol{G}_1$ and $\boldsymbol{G}_3$ with expressions (8.57) and (8.58), gives the propagation equation:

$$\left(-\boldsymbol{\alpha}\left(\sqrt{g}g^{11}\right)\boldsymbol{\alpha} + \partial_3\left(\sqrt{g}g^{33}\right)\partial_3 + i\boldsymbol{\alpha}\left(\sqrt{g}g^{13}\right)\partial_3 + \partial_3\left(\sqrt{g}g^{31}\right) i\boldsymbol{\alpha} + k^2\left(\sqrt{g}g^{22}\right)\right)\boldsymbol{F}_2(x^3) = 0 \tag{8.59}$$

which is rewritten as a pair of first-order differential equation as:

$$-i\partial_3 \boldsymbol{A}\begin{bmatrix}\boldsymbol{F}_2(x^3)\\ -i\partial_3\boldsymbol{F}_2(x^3)\end{bmatrix} = \boldsymbol{B}\begin{bmatrix}\boldsymbol{F}_2(x^3)\\ -i\partial_3\boldsymbol{F}_2(x^3)\end{bmatrix} \tag{8.60}$$

with:

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{\alpha}\left(\sqrt{g}g^{13}\right) + \left(\sqrt{g}g^{13}\right)\boldsymbol{\alpha} & \left(\sqrt{g}g^{33}\right)\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix} \tag{8.61}$$

$$\boldsymbol{B} = \begin{bmatrix}-\boldsymbol{\alpha}\left(\sqrt{g}g^{11}\right)\boldsymbol{\alpha} + k^2\left(\sqrt{g}g^{22}\right) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix} \tag{8.62}$$

Since the coefficients of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are independent of variable x^3, we may seek vectors $\boldsymbol{F}_2(x^3)$, $\boldsymbol{G}_1(x^3)$,$\boldsymbol{G}_3(x^3)$ under the form:

$$\boldsymbol{F}_2(x^3) = \boldsymbol{F_2}\exp(i\rho x^3) \tag{8.63a}$$

$$-i\partial_3\boldsymbol{F}_2(x^3) = \dot{\boldsymbol{F}}_2\exp(i\rho x^3) \tag{8.63b}$$

$$\boldsymbol{G}_1(x^3) = \boldsymbol{G_1}\exp(i\rho x^3) \tag{8.63c}$$

$$\boldsymbol{G}_3(x^3) = \boldsymbol{G_3}\exp(i\rho x^3) \tag{8.63d}$$

This last step transforms (8.60) into a generalized eigenvalue eigenvector matrix equation:

$$\boldsymbol{A}\rho \begin{bmatrix} \boldsymbol{F}_2 \\ \dot{\boldsymbol{F}}_2 \end{bmatrix} = \boldsymbol{B} \begin{bmatrix} \boldsymbol{F}_2 \\ \dot{\boldsymbol{F}}_2 \end{bmatrix} \tag{8.64}$$

It is then easy to check that (8.64) is the same as (8.29). After $\boldsymbol{F}_2$ is determined, it remains to deduce $\boldsymbol{G}_1$ from (8.57).

8.4.2 "Classical" C-method operator

We call "classical" operator the operator derived by Chandezon in his early work. From (8.50b) and taking into account (8.63) we find an expression for $\boldsymbol{G}_3$ as follows:

$$\boldsymbol{G}_3 = \frac{1}{k}\left(\sqrt{g}g^{33}\right)^{-1} i\boldsymbol{\alpha}\boldsymbol{F}_2 - \left(\sqrt{g}g^{33}\right)^{-1}\left(\sqrt{g}g^{31}\right)\boldsymbol{G}_1 \tag{8.65}$$

Substituting $\boldsymbol{G}_3$ in (8.50a) and (8.50c) with the above expression yields:

$$\begin{bmatrix} -\left(\sqrt{g}g^{13}\right)\left(\sqrt{g}g^{33}\right)^{-1}\boldsymbol{\alpha} & ik\left(\left(\sqrt{g}g^{11}\right) - \left(\sqrt{g}g^{13}\right)\left(\sqrt{g}g^{33}\right)^{-1}\left(\sqrt{g}g^{31}\right)\right) \\ -ik\left(\left(\sqrt{g}g^{22}\right) - \frac{1}{k^2}\boldsymbol{\alpha}\left(\sqrt{g}g^{33}\right)^{-1}\boldsymbol{\alpha}\right) & -\boldsymbol{\alpha}\left(\sqrt{g}g^{33}\right)^{-1}\left(\sqrt{g}g^{31}\right) \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_2 \\ \boldsymbol{G}_1 \end{bmatrix} = \rho \begin{bmatrix} \boldsymbol{F}_2 \\ \boldsymbol{G}_1 \end{bmatrix} \tag{8.66}$$

8.5 Multilayer grating

The extension of C-method to multilayer gratings is straightforward provided the interfaces which separate the layers share the same periodicity. It is just a generalization of the theory of planar stratified media. As a canonical case, let us consider a layer made of isotropic homogeneous media limited on the top by surface $z = a_j(x^1)$ and on the bottom by surface $z = a_{j+1}(x^1) = a_j(x^1) - t_j(x^1)$. When $t_j(x^1)$ is constant the two surfaces are parallel to each other. In homogeneous media the field is a superposition of forward and backward waves. The only places where coupling occurs are the interfaces. Thus, we have to describe two different phenomena: on the one hand scattering at the interfaces and on the other hand propagation or attenuation in the layer. To summarize we assimilate an interface to a $4N$-port local network ($N = 4M+1$, M being the truncation number) and a layer to a multi-channel pipe connecting the $2N$-ports of its input network and output network [24]. We have already defined interface scattering matrices which are local matrices in the sense they depend on the profile. In other words,in the context of C-method they depend on the coordinate system. Thus, for a layer bounded by two non parallel surfaces, we have to solve two eigenvalue problems for each surface which allows to calculate interface matrices $\boldsymbol{S}_{a_j}$ and $\boldsymbol{S}_{a_{j+1}}$. It remains to define and to calculate layer scattering matrices. Although two cases have to be considered according to whether the layer separates two identical surfaces or not, the line of reasoning is the same. As already mentioned, we have two coordinate systems such that $z = x_j^3 + a_j(x^1)$ and $z = x_{j+1}^3 + a_{j+1}(x^1)$. They are linked by the following relation:

$$x_j^3 = x_{j+1}^3 + a_{j+1}(x^1) - a_j(x^1) = x_{j+1}^3 - t_j(x^1) \tag{8.67}$$

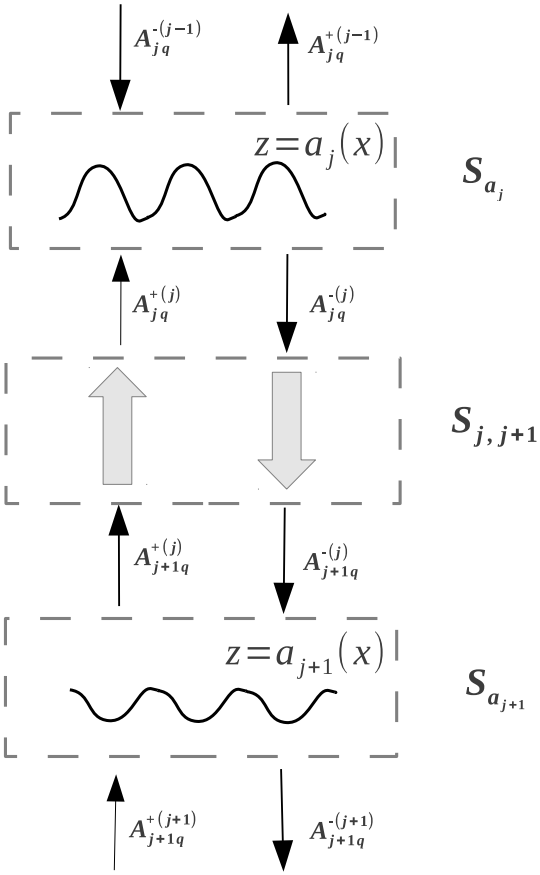


Figure 8.3: Schematic representation of diffraction at two surfaces separated by a layer

.

In medium j, located in between surfaces $z = a_j(x^1)$ and $z = a_{j+1}(x^1)$, we may express the linear combination of forward and backard waves with coordinate $x^3_j = 0$ as local origin (that is $z = a_j(x^1)$ and write:

$$\mathscr{F}^{(j)}_{a_j}(x^3_j, x^1) = \sum_q A^{(j)+}_{j,q} \exp(i\rho^{(j)+}_{a_j,q} x^3_j) F^{(j)+}_{a_j,q}(x^1) + \sum_q A^{(j)-}_{j,q} \exp(i\rho^{(j)-}_{a_j,q} x^3_j)) F^{(j)-}_{a_j,q}(x^1) \quad (8.68)$$

In the same medium j we may also choose $x^3_{j+1} = 0$ as local origin (that is $z = a_{j+1}(x^1)$) which gives:

$$\mathscr{F}^{(j)}_{a_{j+1}}(x^3_{j+1}, x^1)) = \sum_q A^{(j)+}_{j+1,q} \exp(i\rho^{(j)+}_{a_{j+1},q} x^3_{j+1}) F^{(j)+}_{a_{j+1},q}(x^1) + \sum_q A^{(j)-}_{j+1,q} \exp(i\rho^{(j)-}_{a_{j+1},q}(x^3_{j+1}) F^{(j)-}_{a_{j+1},q}(x^1) \quad (8.69)$$

The layer is considered as a $4N - ports$ which connects input waves $\mathscr{F}^{(j)-}_{a_j}$ and $\mathscr{F}^{(j)+}_{a_{j+1}}$ to output waves $\mathscr{F}^{(j)+}_{a_j}$ and $\mathscr{F}^{(j)-}_{a_{j+1}}$, hence the definition of the layer S matrix:

$$\begin{bmatrix} A^{(j)+}_{j,q} \\ A^{(j)-}_{j+1,q} \end{bmatrix} = \boldsymbol{S}_{j,j+1} \begin{bmatrix} A^{(j)-}_{j,q} \\ A^{(j)+}_{j+1,q} \end{bmatrix} \quad (8.70)$$

At the input of the layer, that is at $x^3_j = 0$, the outgoing waves correspond to the incoming wave of the output plane:

$$\mathscr{F}^{(j)+}_{a_j}(x^3_j = 0) = \mathscr{F}^{(j)+}_{a_{j+1}}(x^3_{j+1} = t_j(x^1)) \quad (8.71)$$

Similarly, at the output of the layer, that is $x^3_{j+1} = 0$, the outgoing waves correspond to the incoming waves of the input plane:

$$\mathscr{F}^{(j)-}_{a_{j+1}}(x^3_{j+1} = 0) = \mathscr{F}^{(j)-}_{a_j}(x^3_j = -t_j(x^1) \tag{8.72}$$

At this stage, we infer that layer S matrix looks like:

$$\boldsymbol{S}_{j,j+1} = \begin{bmatrix} 0 & \boldsymbol{P}^{(j)+} \\ \boldsymbol{P}^{(j)-} & 0 \end{bmatrix} \tag{8.73}$$

The sought sub-matrices $\boldsymbol{P}^{(j)+}$, $\boldsymbol{P}^{(j)-}$ depend on whether the layer faces are parallel or not.

8.5.1 Layer with non parallel faces

Consider equation (8.75) and write it in terms of the eigenvectors of both coordinate systems:

$$\sum_m \sum_q A^{(j)+}_{j,q} F^{(j)+}_{a_j,mq} exp(i\alpha_m x^1) = \sum_m \sum_q A^{(j)+}_{j+1,q} \exp(i\rho^+_{a_j+1,q} t_j(x^1)) F^{(j)+}_{a_{j+1},q} \exp(i\alpha_m x^1) \tag{8.74}$$

The left hand side purely consists of a linear combination of eigenvectors expanded onto the $\exp(i\alpha_m x^1)$ basis whereas the right hand side consists of a linear combination of eigenvectors each of which being multiplied by a periodic functions of the x^1 variable. In order to get a matrix relation between the $A^{(j)+}_{j,q}$ and the $A^{(j)+}_{j+1,q}$, we project (8.74) onto $\exp(i\alpha_m x^1)$. We get:

$$\sum_m \sum_q A^{(j)+}_{j,q} F^{(j)+}_{a_j,mq} = \sum_m \sum_q A^{(j)+}_{j+1,q} \tilde{F}^{(j)+}_{a_{j+1},q} \exp(i\alpha_m x^1) \tag{8.75}$$

with:

$$\tilde{F}^{(j)+}_{a_{j+1}mq} = \frac{1}{d^1} \int_0^{d^1} \left(\sum_l F^{(j)+}_{a_{j+1},lq} \exp(i\alpha_l x^1) \right) \exp(i\rho^{(j)+}_{a_j+1,q} t_j(x^1)) \exp(-i\alpha_m x^1) dx^1 \tag{8.76}$$

Then, the $\boldsymbol{P}^{(j)+}$ matrix is readily obtained as

$$\boldsymbol{P}^{(j)+} = \left(F^{(j)+}_{a_j} \right)^{-1} \tilde{\boldsymbol{F}}^{(j)+}_{a_{j+1}} \tag{8.77}$$

where $\boldsymbol{F}^{(j)+}_{a_j}$ (respectively $\tilde{\boldsymbol{F}}^{(j)+}_{a_{j+1}}$) is the matrix formed by juxtaposition of vectors $F^{(j)+}_{a_j,q}$ (respectively $\tilde{F}^{(j)+}_{a_{j+1}q}$. Similarly we have:

$$\sum_m \sum_q A^{(j)-}_{j+1,q} F^{(j)-}_{a_{j+1},mq} \exp(i\alpha_m x^1) = \sum_m \sum_q A^{(j)-}_{j,q} \exp(i\rho^{(j)-}_{a_j,q} t_j(x^1)) F^{(j)-}_{a_j,q} \exp(i\alpha_m x^1) \tag{8.78}$$

and

$$\sum_m \sum_q A^{-(j)}_{j+1,q} F^{(j)-}_{a_{j+1},mq} = \sum_m \sum_q A^{(j)-}_{j,q} \tilde{F}^{(j)-}_{a_j,q} \tag{8.79}$$

with:

$$\tilde{F}^{(j)-}_{a_jmq} = \frac{1}{d^1} \int_0^{d^1} \left(\sum_l F^{(j)-}_{a_j,lq} \exp(i\alpha_l x^1) \right) \exp(-i\rho^{(j)-}_{a_j+1,q} t_j(x^1)) \exp(-i\alpha_m x^1) dx^1 \tag{8.80}$$

from which we derive:

$$\boldsymbol{P}^{(j)-} = \left(F^{(j)-}_{a_{j+1}} \right)^{-1} \tilde{\boldsymbol{F}}^{(j)-}_{a_j} \tag{8.81}$$

8.5.2 Layer with parallel faces

In that case, the two coordinate systems are identical and $t_j(x^1)$ is a constant. Equations (8.72) and (8.75) reduce to:

$$\sum_m \sum_q A_{j,q}^{(j)+} F_{a_j,mq}^{(j)+} \exp(i\alpha_m x^1) = \sum_m \sum_q A_{j+1,q}^{+(j)} \exp(i\rho_{a_j,q}^{(j)+} t_j) F_{a_j,mq}^{(j)+} \exp(i\alpha_m x^1) \tag{8.82}$$

$$\sum_m \sum_q A_{j+1,q}^{(j)-} F_{a_j,q}^{(j)+} = \sum_m \sum_q A_{j,q}^{(j)-} \exp(-i\rho_{a_j,q}^{(j)-} t_j) F_{a_j,mq}^{(j)-} \exp(i\alpha_m x^1) \tag{8.83}$$

from which we easily deduce:

$$A_{j,q}^{(j)+} = A_{j+1,q}^{(j)+} \exp(i\rho_{a_j,q}^{(j)+} t_j) \text{ or } \boldsymbol{P}^{(j)+} = \text{diag}\left(\exp(i\rho_{a_j,q}^{(j)+} t_j)\right) \tag{8.84}$$

$$A_{j+1,q}^{(j)-} = A_{j,q}^{(j)-} \exp(-i\rho_{a_j,q}^{(j)-} t_j) \text{ or } \boldsymbol{P}^{(j)-} = \text{diag}\left(\exp(-i\rho_{a_j,q}^{(j)-} t_j)\right) \tag{8.85}$$

It should be noted that when $\rho_{a_j,q}^{(j)+}$ (respectively $\rho_{a_j,q}^{(j)-}$) is complex valued, its imaginary part is negative (respectively positive) . Since t_j is positive, exponential functions $\exp(\pm i\rho_{a_j,q}^{(j)\pm} t_j)$ associated to complex eigenvalues always decay when the layer thickness increases.

8.5.3 Combination of S matrices

The final step for analysing reflection and transmission by a layer is to combine the two interfaces S matrix and the layer S matrix. The tool for doing this is the Redheffer star product which gives the composition rules of two cascaded S matrices [26]. Consider two S matrices and partition them into four blocks:

$$\boldsymbol{S}_1 = \begin{bmatrix} \boldsymbol{S}_1^{11} & \boldsymbol{S}_1^{12} \\ \boldsymbol{S}_1^{21} & \boldsymbol{S}_1^{22} \end{bmatrix} \quad \boldsymbol{S}_2 = \begin{bmatrix} \boldsymbol{S}_2^{11} & \boldsymbol{S}_2^{12} \\ \boldsymbol{S}_2^{21} & \boldsymbol{S}_2^{22} \end{bmatrix} \tag{8.86}$$

The star product $*$ is defined by:

$$\boldsymbol{S} = \boldsymbol{S_1} * \boldsymbol{S_2} \tag{8.87}$$

$$\boldsymbol{S}^{11} = \boldsymbol{S}_1^{11} + \boldsymbol{S}_1^{12} \left(\boldsymbol{I} - \boldsymbol{S}_2^{11} \boldsymbol{S}_1^{22}\right)^{-1} \boldsymbol{S}_2^{11} \times \boldsymbol{S}_1^{21} \tag{8.88}$$

$$\boldsymbol{S}^{12} = \boldsymbol{S}_1^{12} \times \left(\boldsymbol{I} - \boldsymbol{S}_2^{11} \boldsymbol{S}_1^{22}\right)^{-1} \times \boldsymbol{S}_2^{12} \tag{8.89}$$

$$\boldsymbol{S}^{21} = \boldsymbol{S}_2^{21} \times \left(\boldsymbol{I} - \boldsymbol{S}_1^{22} \boldsymbol{S}_2^{11}\right)^{-1} \times \boldsymbol{S}_1^{21} \tag{8.90}$$

$$\boldsymbol{S}^{22} = \boldsymbol{S}_2^{22} + \boldsymbol{S}_2^{21} \times \left(\boldsymbol{I} - \boldsymbol{S}_1^{22} \boldsymbol{S}_2^{11}\right)^{-1} \times \boldsymbol{S}_1^{22} \times \boldsymbol{S}_2^{12} \tag{8.91}$$

where $\boldsymbol{I}$ is the identity matrix. The combined $\boldsymbol{S}$ matrix of the top and bottom interfaces and of the layer is given by:

$$\boldsymbol{S} = \boldsymbol{S}_{a_j} * \boldsymbol{S}_{j,j+1} * \boldsymbol{S}_{a_{j+1}} = \left(\boldsymbol{S}_{a_j} * \boldsymbol{S}_{j,j+1}\right) * \boldsymbol{S}_{a_{j+1}} = \boldsymbol{S}_{a_j} * \left(\boldsymbol{S}_{j,j+1} * \boldsymbol{S}_{a_{j+1}}\right) \tag{8.92}$$

and finally, it turns out that

$$\boldsymbol{S}^{11} = \boldsymbol{S}_{a_j}^{11} + \boldsymbol{S}_{a_j}^{12} \boldsymbol{P}(j)^{+} \boldsymbol{U}_2 \boldsymbol{P}^{(j)-} \boldsymbol{S}_{a_{j+1}}^{11} \tag{8.93}$$

$$\boldsymbol{S}^{12} = \boldsymbol{S}_{a_j}^{12} \boldsymbol{P}^{(j)+} \boldsymbol{U}_2 \boldsymbol{S}_{a_{j+1}}^{12} \tag{8.94}$$

$$\boldsymbol{S}^{21} = \boldsymbol{S}_{a_{j+1}}^{21} \boldsymbol{U}_1 \boldsymbol{P}^{(j)-} \boldsymbol{S}_{a_j}^{21} \tag{8.95}$$

$$\boldsymbol{S}^{22} = \boldsymbol{S}_{a_{j+1}}^{22} + \boldsymbol{S}_{a_{j+1}}^{21} \boldsymbol{P}^{(j)-} \boldsymbol{U}_1 \boldsymbol{P}^{(j)+} \boldsymbol{S}_{a_j}^{22} \tag{8.96}$$

where

$$\boldsymbol{U}_1 = \left(\boldsymbol{I} - \boldsymbol{S}_{a_j}^{22}\boldsymbol{P}^{(j)+}\boldsymbol{S}_{a_{j+1}}^{11}\boldsymbol{P}^{(j)-}\right)^{-1} \boldsymbol{U}_2 = \left(\boldsymbol{I} - \boldsymbol{S}_{a_{j+1}}^{11}\boldsymbol{P}^{(j)-}\boldsymbol{S}_{a_j}^{22}\boldsymbol{P}^{(j)+}\right)^{-1}$$

8.6 Extensions of C Method

The key idea of C-method as applied to diffraction by surface-relief gratings is to map the surface of the grating to a plane. Until now, we have only described profiles under the form $z = a(x)$. However, in Cartesian coordinates (x,y,z), a cylindrical surface whose generating line is parallel to the *Oy* axis may be described by the parametric equations:

$$x = f(x^1) \;\; z = g(x^1) \tag{8.97}$$

where f and g are two continuous functions. Now consider the following relations:

$$x = f(x^1) + c_1 x^3 \;\; y = x^2 \;\; z = g(x^1) + x^3 \tag{8.98}$$

where c_1 is a real constant. They define an additive change of coordinates whose metric tensor is given by:

$$g_{ij} = \begin{bmatrix} \partial_1^2 f + \partial_1^2 g & 0 & c_1\partial_1 f + \partial_1 g \\ 0 & 1 & 0 \\ c_1\partial_1 f + \partial_1 g & 0 & 1 + \partial_1^2 g \end{bmatrix} \tag{8.99}$$

Actually, the above matrix corresponds to a change of coordinates provided the Jacobian determinant J of the transformation does not go to zero.

$$J = \begin{vmatrix} \partial_1 x & \partial_3 x \\ \partial_1 z & \partial_3 z \end{vmatrix} = \begin{vmatrix} \partial_1 f & c_1 \\ \partial_1 g & 1 \end{vmatrix} = \partial_1 f - c_1\partial_1 g \tag{8.100}$$

More over, the metric tensor is independent of coordinate x^3 which means there exists a translation symmetry along x^3 axis. Hence Equations(8.98) define in a general way translation coordinate systems which allow to solve new classes of problems.

8.6.1 Oblique transformations

In Cartesian coordinates, usual coordinates lines of a plane are two straight lines orthogonal to each other. One can also imagine having straight lines which make an angle different from $\pi/2$. Consider the straight line Δ given by $z = \tan(\phi)x$ and let us call ϕ the obliquity angle. The following sets of relations define a coordinate system (x^1, x^3) in which lines parallel to Δ are coordinate lines $x^1 = constant$ and lines $x^3 =$ constant remain parallel to *Ox*

$$\begin{aligned} x &= x^1 + \frac{1}{\tan\phi} x^3 \\ z &= x^3 \end{aligned} \tag{8.101}$$

Such oblique transformation allow to model an extended class of surface shapes which would otherwise be numerically inefficient (very blazed gratings) or even impossible like overhanging gratings. As an illustrative example, consider in the coordinate system (x^1, x^3) the symmetric triangular function.

$$t(x^1) = \begin{cases} 2x^1 & 0 < x^1 < .5 \\ 2(1-x^1) & .5 < x^1 < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{8.102}$$

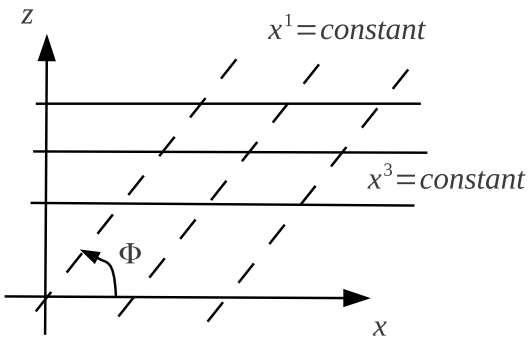


Figure 8.4: Coordinate system in which coordinate lines are parallel to Δ and to Ox axis

Using an oblique transformation one gets:

$$\begin{aligned} x &= x^1 + \frac{1}{\tan\phi}(x^3 + t(x^1)) \\ z &= x^3 + t(x^1) \end{aligned} \tag{8.103}$$

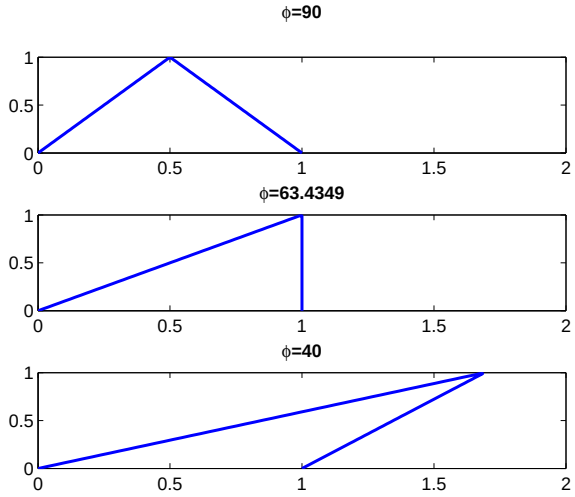


Figure 8.5: Echelette grating in three different oblique coordinate systems

Figure(8.5) shows three typical grating surfaces obtained with (8.103) and with $\phi = 90$, 63.4349 and 40 respectively, the latter demonstrating the extreme overhanging forms possible for small ϕ without the double value problem implicit with Cartesian coordinates.

8.6.2 Stretched coordinates

The essence of C-method is to choose a coordinate system that facilitate the solution of a given problem. Oblique transformations are a typical example of the usefulness of this technique. Indeed they provide an easy and elegant way to handle gratings with one vertical facet and also overhanging gratings. Similarly, we have believed for a long time that sharp edges were an intrinsic limitation of the C method. Actually, it turns out that transformations which stretch coordinates around the edges overcome the problem. With C-Method, the solution of Maxwell's equations is reduced to the solution of an algebraic eigenvalue problem in discrete Fourier

space. The derivation of the matrix operator involves two steps: (1) The electromagnetic field is expanded into Floquet–Fourier series, and (2) the derivative of the grating profile function is expanded into Fourier series. When the latter function is discontinuous, the Fourier method is known to converge slowly.This weakness remains even when the correct Fourier factorization of products of discontinuous periodic functions, as given by Li, is applied. The reason for slow convergence is that the spatial resolution of the Fourier expansion remains uniform within a grating period whatever the grating profile function may be. On the contrary, stretched coordinates allow a mapping of space that increases spatial resolution around the discontinuities of the derivative of the profile function. For this reason the technique is known as adaptive spatial resolution.

8.6.3 Parametric C-method

Whether for mandatory reasons as is the case for overhanging gratings or simply to improve convergence speed, the most general representation of a one dimensional profile happens to be a parametric one. Adding an additional degree of freedom with an obliquity angle, a class of translation coordinate systems has the form given by (8.98). Due to the translational symmetry along vector $\boldsymbol{e}_3 = c_1\boldsymbol{e}_x + \boldsymbol{e}_z$, a numerical solution in terms of eigenvectors and eigenvalues is possible. Equations (8.98) describe a coordinate system where coordinates lines $x^3 =$ constant coincide with functions which are periodic with period d^1 along direction Ox. Compared to the non-oblique coordinate system, the period d^1 and the direction of periodicity remain unaffected by the introduction of parameter c_1. Thus, assuming an incident plane wave vector $\boldsymbol{k}$ such that $\boldsymbol{k}.\boldsymbol{e}_x = \alpha_0$ the x^1 dependence is of the form $\exp(i\alpha_m x^1)$ with $\alpha_m = \alpha_0 + m\dfrac{2\pi}{d^1}$. So we have all the ingredients to determine the matrix from which eigenvectors and eigenvalues will be sought. In Fourier space, the matrices associated to the elements of the metric tensor are:

$$\begin{aligned}
\left(\sqrt{g}g^{11}\right) &= (c_1^2+1)\left[\dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}\right]^{-1} \\
\left(\sqrt{g}g^{13}\right) &= \left[\dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}\right]^{-1}\left[c_1\dot{\boldsymbol{f}} + \dot{\boldsymbol{g}}\right] \\
\left(\sqrt{g}g^{31}\right) &= \left[c_1\dot{\boldsymbol{f}} + \dot{\boldsymbol{g}}\right]\left[\dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}\right]^{-1} \\
\left(\sqrt{g}g^{33}\right) &= \left[\dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}\right] + (c_1^2+1)\left[c_1\dot{\boldsymbol{f}} + \dot{\boldsymbol{g}}\right]\left[\dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}\right]^{-1}\left[c_1\dot{\boldsymbol{f}} + \dot{\boldsymbol{g}}\right] \\
\left(\sqrt{g}g^{22}\right) &= \dot{\boldsymbol{f}} - c_1\dot{\boldsymbol{g}}
\end{aligned} \tag{8.104}$$

where $\dot{f}$ and $\dot{g}$ designates the toeplitz matrices formed by the elements of the Fourier series of $\partial_1 f$ and $\partial_1 g$ respectively.

8.6.4 Plane waves and parametric C-method

More over since the physics remains the same compared to non-oblique translation coordinate systems, eigenvectors separate into forward and backward waves as was already the case:

$$\mathscr{F}(x^1,x^3) = \sum_q A_q^{\pm} F_q^{\pm}(x^1) exp(i\rho_q^{\pm} x^3) \tag{8.105}$$

As in the classical translation coordinate system, we substitute the computed propagative forward and backward eigenvectors with the corresponding transformed plane waves . Consider

plane waves

$$\exp(i\alpha_n x)\exp(\pm i\gamma_n z)$$

such that $\pm\gamma_n \in R$. Taking into account (8.98),their expression in oblique coordinates is:

$$\exp\left(i(\alpha_n c_1 \pm \gamma_n)x^3\right)\exp\left(i(\alpha_n f(x^1) \pm \gamma_n g(x^1)\right) \tag{8.106}$$

hence the following correspondences between propagative waves in Cartesian coordinates and their computed counterparts in oblique coordinates:

$$\rho^{\pm,(M)}_{(p),n} \longleftrightarrow (\pm\gamma_n + \alpha_n c_1)\,; \quad F^{\pm}_{(p),n}(x^1) \longleftrightarrow \exp\left(i\alpha_n f(x^1) \pm i\gamma_n(g(x^1)\right) \tag{8.107}$$

We have added an extra subscript (p) and a superscript (M) to indicate that we only care about the above correspondence for propagative waves and that ρ depends on the truncation number.

8.6.5 Illustrative example

Consider a right angled triangular profile whose base is aligned on Ox. Other parameters are period d^1 and height h. It is illuminated by a plane wave inclined at θ to the Oz axis. In the context of C-method we ask ourselves which coordinate system choosing for modelling diffraction by such a grating. Here the main difficulty comes from the vertical facet located at $x = d^1$. The operator associated with C-method involves the derivative of the profile function. With a description of the profile by a function of the kind $z = a(x)$, the derivative is constant and everything happens as if the vertical did not exist. Should the vertical be replaced by a very sloping facet, then a highly located and large discontinuity in the derivative would appear. None of the situation is satisfactory. An easy way to overcome the problem consists in introducing an oblique coordinate system in which the vertical is transformed into a straight line with a "reasonable" slope. Actually, doing so amounts to parametrizing the profile in the Cartesian coordinate system.

8.6.5.1 Obliquity angle and parametrization of the profile

Since one of the facets of the grating is vertical, an inclined coordinate system is needed. On the one hand, the parameter c_1 is linked to the obliquity angle ϕ by $c_1 = 1/\tan\phi$ and on the other hand, according to (8.100) it should satisfy the constraint $1 - c_1\partial_x a > 0$. Hence, in principle ϕ may be any angle such that $\tan\phi < h/d$ Let t_1 be $\tan(\phi)$. On the first facet we have:

$$x = x^1 + \frac{1}{t_1}y, \qquad y = \frac{h}{d}x \tag{8.108}$$

and on the second one

$$x = d, \qquad d = x^1 + \frac{1}{t_1}y \tag{8.109}$$

Thus the parametrization of the profile is:

$$\begin{aligned} f(x^1) &= \frac{d}{d - t_1 h}x^1 \quad g(x^1) = \frac{h}{d - t_1 h}x^1 \quad \text{if } x^1 \leqslant x_0^1 \\ f(x^1) &= d \qquad\qquad\quad g(x^1) = \frac{1}{t_1}(x^1 - d) \quad \text{if } x^1 > x_0^1 \end{aligned} \tag{8.110}$$

with :

$$x_0^1 = d\left(1 - \frac{1}{t_1}\frac{h}{d}\right) \tag{8.111}$$

Now that we have parametrized the profile, it remains to define a translation direction. The direction which served to parametrize the profile is a natural choice although not mandatory. Once more, the only constraint is that the tangent of the chosen obliquity angle is smaller than h/d.

8.6.5.2 Stretched coordinates and parametrization of the profile

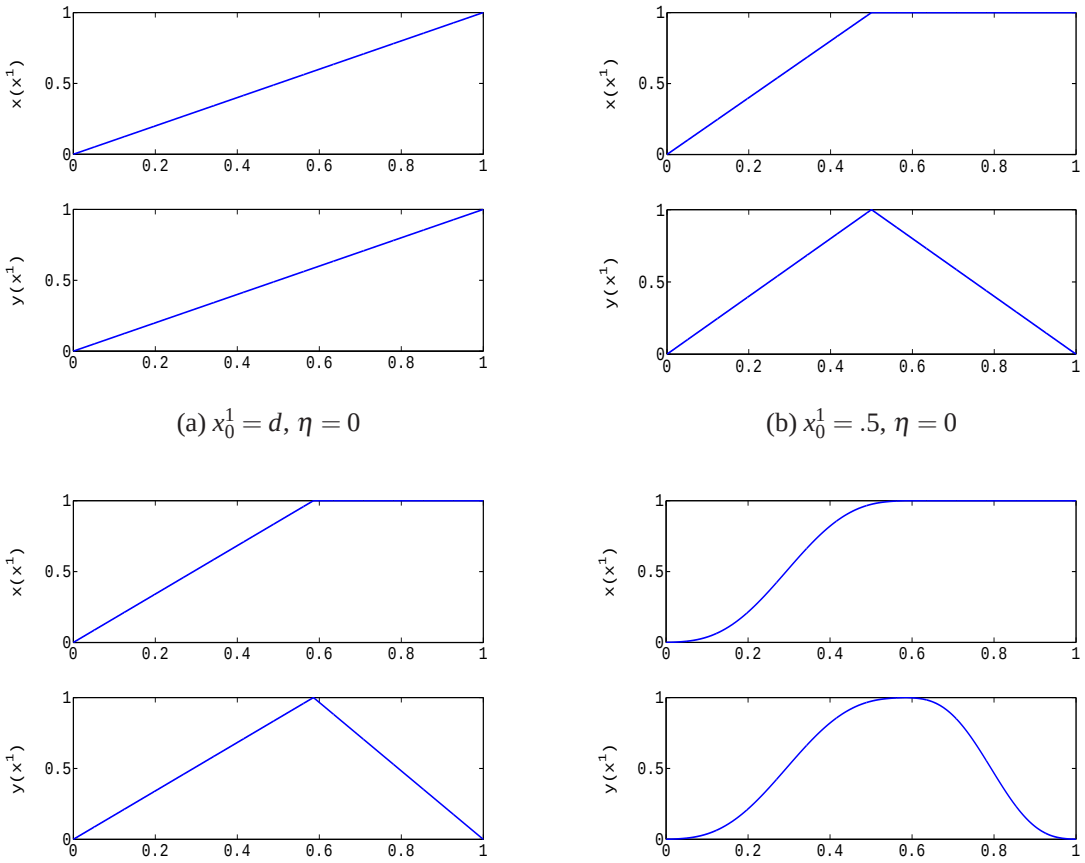
At $x^1 = 0$ and $x^1 = x_0^1$, the parametric functions $f(x^1)$ and $g(x^1)$ have jumps which can be reduced if one introduces an additional change of coordinates aimed at increasing spatial resolution around these points. Let x^1 be a function of a new variable u: $x^1 = s(u)$. The chain rule for derivative gives :

$$\partial_u x = \partial_1 f(x^1(s(u))\partial_u s, \qquad \partial_u y = \partial_1 g(x^1(s(u))\partial_u s \tag{8.112}$$

Compared to the initial parametrization, spatial resolution is modulated by the multiplicative factor $\partial_u s$. The smaller the latter, the higher the spatial resolution. A possible stretching function is as follows :

$$s(u) = \begin{cases} u - \dfrac{\eta x_0}{2\pi}\sin\left(\dfrac{2\pi u}{x_0}\right) & \text{if } 0 \leqslant u < x_0^1 \\ (u - x_0) - \dfrac{\eta(d^1 - x_0^1)}{2\pi}\sin\left(\dfrac{2\pi(u-x_0)}{d^1 - x_0^1}\right) & \text{if } x_0^1 \leqslant u < d^1 \end{cases} \tag{8.113}$$

The parameter η between zero and one controls the density of coordinate lines around the transition points. It allows to stretch space thinner where discontinuities of coefficients in Maxwell's equations occur. The larger η, the smaller $\partial_u s$ and thus the higher the spatial resolution. In principle the parameter η does not have to reach one because, in that case, the Jacobian would be zero. Figure (8.6) shows four possible parametrization of the considered right angle triangle. Case (a) corresponds to the usual representation $z = a(x)$.

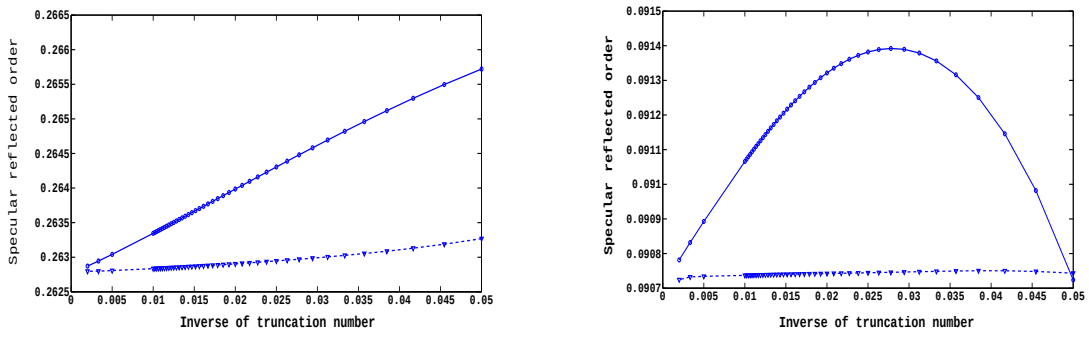


(a) $x_0^1 = d$, $\eta = 0$ (b) $x_0^1 = .5$, $\eta = 0$

(c) $x_0^1 = .5858$, $\eta = 0$ (d) $x_0^1 .5858$ $\eta = .95$

Figure 8.6: Various parametric representations of a right angle triangular profile

Finally, figure (8.7) shows the speed of convergence of the specular reflected order for a perfectly conducting right angle triangular profile for two different parametrizations. It has to be emphasized that modelling this kind of profile is out of reach for the "classical" C-method since it has a vertical facet.



(a) TE polarization, (b) TM polarization

Figure 8.7: Comparison of speed of convergence for two parametric representation of a right angle triangular profile. Full line: $x_0^1 = .5$, $\eta = 0$, dashed line: $x_0^1 = .4$, $\eta = .9$. Other parameters are: $\theta = 25°$, $\lambda = 1$,$h = d^1 = 1$

Appendix 8.A: Curvilinear Coordinates

In Cartesian coordinates we deal with three mutually perpendicular families of planes: x=constant, y=constant, z=constant. Imagine that we superimpose on this system three other families of surfaces. We may reference any variable point M by the intersection of three planes in Cartesian coordinates , ie by the triplet (x,y,z) or as the intersection of the three surfaces that form our new, curvilinear coordinates. Describing the curvilinear coordinates surfaces by $x^1 = constant$,$x^2 = constant$,$x^3 = constant$ we may identify our point by the triplet x,y,z as well as by x^1,x^2,x^3. This means that in principle we may define a curvilinear coordinate system from the Cartesian system (x,y,z) by:

$$x = x^{1'} = x^{1'}(x^1,x^2,x^3), y = x^{2'} = x^{2'}(x^1,x^2,x^3), z = \vec{x^3} = x^{3'}(x^1,x^2,x^3) \tag{8.114}$$

or by the inverse relations

$$x^1 = x^1(x^{1'},x^{2'},x^{3'}), x^2 = x^2(x^{1'},x^{2'},x^{3'}), x^3 = x^3(x^{1'},x^{2'},x^{3'}) \tag{8.115}$$

$x^{1'},x^{2'},x^{3'}$ respectively x^1,x^2,x^3 are regarded as independent and continuously differentiable functions of x^1, x^2 and x^3 respectively $x^{1'}$,$x^{2'}$,$x^{3'}$. let M denote a variable point referenced by the rectangular coordinates(x,y,z).At M the so-called natural referential$(M,\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3)$is defined by the the following basis vectors:

$$\boldsymbol{e}_\alpha = \sum_{\beta'=1}^{\beta'=3} \frac{\partial x^{\beta'}}{\partial x^\alpha} \boldsymbol{e}_{\beta'} \tag{8.116}$$

with $e_{1'} = e_x$, $e_{2'} = e_y$, $e_{3'} = e_z$, e_x, e_y and e_z being the unit vectors of an orthogonal Cartesian referential. In a similar way we may write

$$\boldsymbol{e}_{\alpha'} = \sum_{\beta=1}^{\beta=3} \frac{\partial x^{\beta}}{\partial x^{\alpha'}} \boldsymbol{e}_{\beta} \tag{8.117}$$

Moreover introducing $\Lambda_\alpha^{\beta'} = \dfrac{\partial x^{\beta'}}{\partial x^\alpha}$ and Einsteins' summation convention Eq(8.116) and Eq(8.117) write:

$$\boldsymbol{e}_\alpha = \Lambda_\alpha^{\beta'} \boldsymbol{e}_{\beta'} \qquad \boldsymbol{e}_\alpha = \Lambda_\alpha^{\beta} \boldsymbol{e}_{\beta'} \tag{8.118}$$

vectors $\boldsymbol{e}_\alpha$ are tangent vectors along coordinate curve x^α. The matrix formed by the coefficient$\Lambda_\alpha^{\beta'}$ is the Jacobian matrix $\boldsymbol{J}$ of the change of coordinates. Since functions $x^{1'}$, $x^{2'}$,$x^{3'}$ are independent $\boldsymbol{J}$ is inverible and its inverse is formed by the coefficients $\Lambda_{\alpha'}^{\beta}$

$$\boldsymbol{J} = \begin{bmatrix} \Lambda_1^{1'} & \Lambda_1^{2'} & \Lambda_1^{3'} \\ \Lambda_2^{1'} & \Lambda_2^{2'} & \Lambda_2^{3'} \\ \Lambda_3^{1'} & \Lambda_3^{2'} & \Lambda_3^{3'} \end{bmatrix} \qquad \boldsymbol{J}^{-1} = \begin{bmatrix} \Lambda_{1'}^{1} & \Lambda_{1'}^{2} & \Lambda_{1'}^{3} \\ \Lambda_{2'}^{1} & \Lambda_{2'}^{2} & \Lambda_{2'}^{3} \\ \Lambda_{3'}^{1} & \Lambda_{3'}^{2} & \Lambda_{3'}^{3} \end{bmatrix} \tag{8.119}$$

One can also define basis vectors $\boldsymbol{e}^\alpha$ that are normal to coordinate surfaces $x^\alpha = \text{constant}$ by

$$\boldsymbol{e}^\alpha = \frac{\partial x^\alpha}{\partial x^{\alpha'}} \boldsymbol{e}^{\alpha'}, \qquad \text{with } \boldsymbol{e}^{\alpha'} = \boldsymbol{e}_{\alpha'} \tag{8.120}$$

The vectors $\boldsymbol{e}_\alpha$ and $\boldsymbol{e}^\beta$ form a set of reciprocal basis with $\boldsymbol{e}_\alpha.\boldsymbol{e}^\beta = \delta_\alpha^\beta$, where δ_α^β is the Kronecker delta. The representation of any vector $\boldsymbol{A}$ in one of these bases is:

$$\boldsymbol{A} = A^\alpha \boldsymbol{e}_\alpha = A_\alpha \boldsymbol{e}^\alpha \tag{8.121}$$

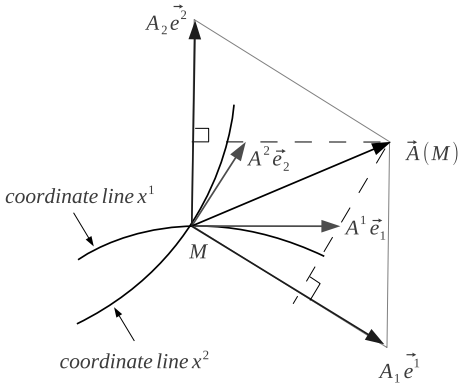


Figure 8.8: Curvilinear Coordinates:covariant and contravariant components of a vector in a plane

The A^α and the A_α are the contravariant components and the covariant components of vector $\boldsymbol{A}$ respectively. The nullity of a component of vector $\boldsymbol{A}$ may be geometrically interpreted as follows:
$A_\alpha = 0$: $\boldsymbol{A}$ is orthogonal to the tangent at point M to the coordinate line x^α
$A^\alpha = 0$: $\boldsymbol{A}$ belongs to the tangential plane at point M to coordinate surface x^α
In normalized orthogonal Cartesian coordinates differentiate contravariant and covariant components of a vector is generally not necessary. The Jacobian matrix allows to express the Cartesian components $A^{\alpha'} = A_{\alpha'}$ of vector $\boldsymbol{A}$ in terms of its local contravariant components $A^\alpha = \boldsymbol{A}.\boldsymbol{e}^\alpha$ or covariants components $A_\alpha = \boldsymbol{A}.\boldsymbol{e}_\alpha$:

$$A^{\alpha'} = A_{\alpha'} = \Lambda^{\alpha'}_{\alpha} A^\alpha \qquad \text{or} \qquad A^\alpha = \Lambda^{\alpha}_{\alpha'} A^{\alpha'} \tag{8.122}$$

$$A_{\alpha'} = \Lambda^{\alpha}_{\alpha'} A_\alpha \qquad \text{or} \qquad A_\alpha = \Lambda^{\alpha'}_{\alpha} A_{\alpha'} \tag{8.123}$$

The quantities

$$g_{\alpha\beta} = \boldsymbol{e}_\alpha.\boldsymbol{e}_\beta = \Lambda^{\alpha'}_{\alpha} \Lambda^{\beta'}_{\beta} g_{\alpha'\beta'} \tag{8.124}$$

define the metric of the coordinate system. In matrix form we have:

$$\left[g_{\alpha\beta}\right] = \boldsymbol{J}^t \boldsymbol{J} \tag{8.125}$$

and

$$g = \det(\left[g_{\alpha\beta}\right]) = \det(\boldsymbol{J})^2 = \det\left(\Lambda^{\alpha'}_{\beta}\right) \tag{8.126}$$

The $g_{\alpha\beta}$ establish a connexion between the A^α and the A_β

$$A_\alpha = \boldsymbol{e}_\alpha.\left(A^\beta.\boldsymbol{e}_\beta\right) = g_{\alpha\beta} A^\beta \qquad \text{or} \qquad A^\beta = g^{\beta\alpha} A_\alpha \tag{8.127}$$

Appendix 8.B: Transformation of Maxwell's equations

We have seen that the natural referentiel gives the tools to easily maipulate tangential and normal components of a vector field.Therefore, writting boundary conditions at a surface should be

straightforward. We need now to express Maxwell's equation in the new coordinate system. For that purpose, we may follow a tensorial approach or stay at an elementary level and make a simple change of coordinates and components in the usual Maxwell's equations. We present briefly both points of view. A time dependence of the form $\exp(-i\omega t)$ is assumed.

Vectorial approch

Let us start from one of the Maxwell's curl equation written in the Cartesian coordinate system and in an homogeneous medium with permittivity ε and permeability μ

$$\xi^{\alpha'\beta'\gamma'}\partial_{\beta'}H_{\gamma'} = -i\omega\varepsilon E_{\alpha'} \tag{8.128}$$

where $\xi^{\alpha'\beta'\gamma'}$ stands for the Levi-Civita indicator :

$$\xi^{\alpha'\beta'\gamma'} = \begin{cases} 1 \text{ for } \alpha'\beta'\gamma' = 123, 231, 312 \\ -1 \text{ for } \alpha'\beta'\gamma' = 321, 213, 132 \\ 0 \text{ otherwise} \end{cases} \tag{8.129}$$

Then let us change the coordinates: $\partial_{\beta'} = \Lambda^{\alpha}_{\beta'}\partial_{\alpha}$ and the components $H_{\gamma'} = \Lambda^{\beta}_{\gamma'}H_{\beta}$

$$\xi^{\alpha'\beta'\gamma'}\Lambda^{\alpha}_{\beta'}\partial_{\alpha}\left(\Lambda^{\beta}_{\gamma'}H_{\beta}\right) = -i\omega\varepsilon E_{\alpha'} \tag{8.130}$$

The left hand side of the above equation is equal to:

$$\xi^{\alpha'\beta'\gamma'}\Lambda^{\alpha}_{\beta'}\left(\Lambda^{\beta}_{\gamma'}\partial_{\alpha}H_{\beta}\right) + \xi^{\alpha'\beta'\gamma'}\Lambda^{\alpha}_{\beta'}\left(\partial_{\alpha}\Lambda^{\beta}_{\gamma'}\right)H_{\beta} \tag{8.131}$$

on the one hand we have

$$\Lambda^{\alpha}_{\beta'}\partial_{\alpha}\Lambda^{\beta}_{\gamma'} = \partial_{\beta'}\Lambda^{\beta}_{\gamma'} \tag{8.132}$$

and the other hand this term is symmetrical with respect to β' and γ'. Thus by applying the operator $\xi^{\alpha'\beta'\gamma'}$ which is antisymmetric with respect to β' and γ' we obtain 0. Thus, the Maxwell curl equation reduces to:

$$\xi^{\alpha'\beta'\gamma'}\Lambda^{\alpha}_{\beta'}\Lambda^{\beta}_{\gamma'}\partial_{\alpha}H_{\beta} = -i\omega\varepsilon E_{\alpha'} \tag{8.133}$$

let us multiply both sides by $\Lambda^{\gamma}_{\alpha'}$ and make summation on dummy index α'. We obtain:

$$\xi^{\gamma\alpha\beta}\det\left(\Lambda^{\alpha}_{\beta'}\right)\partial_{\alpha}H_{\beta} = -i\omega\varepsilon\Lambda^{\gamma}_{\alpha'}E^{\alpha'} = -i\omega\varepsilon E^{\gamma} = -i\omega\varepsilon g^{\gamma\beta}E_{\beta} \tag{8.134}$$

Finally we get:

$$\xi^{\gamma\alpha\beta}\partial_{\alpha}H_{\beta} = -i\omega\varepsilon\sqrt{g}g^{\gamma\beta}E_{\beta} \tag{8.135}$$

and

$$\xi^{\gamma\alpha\beta}\partial_{\alpha}E_{\beta} = i\omega\mu\sqrt{g}g^{\gamma\beta}H_{\beta} \tag{8.136}$$

setting

$$\mathscr{F}_{\alpha} = E_{\alpha} \qquad \mathscr{G}_{\alpha} = iZH_{\alpha} \qquad \text{with } Z = \sqrt{\frac{\mu}{\varepsilon}} \tag{8.137}$$

we then obtain a set of equations relating the complex amplitudes of the field components where the $\mathscr{F}_{\alpha}$ and the $\mathscr{G}_{\alpha}$ play a fully symmetric role:

$$\begin{aligned} \xi^{\alpha\beta\gamma}\partial_{\beta}\mathscr{F}_{\gamma} &= k\sqrt{g}g^{\alpha\beta}\mathscr{G}_{\alpha} \\ \xi^{\alpha\beta\gamma}\partial_{\beta}\mathscr{G}_{\gamma} &= k\sqrt{g}g^{\alpha\beta}\mathscr{F}_{\alpha} \end{aligned} \tag{8.138}$$

where $k = \omega\sqrt{\mu\varepsilon}$

Appendix 8.C: Summary of tensorial approach

In curvilinear coodinates systems, the Maxwell's equations are based on the tensorial formalism deduced from relativity. If we consider only materials which are stationary with respect to the coordinate system, then the four-dimensional formalism developed by Post can be simplified. Maxwell equations are written :

$$\begin{aligned} &\xi^{\alpha\beta\gamma}\partial_\beta E_\gamma = -\partial_t B^\alpha \\ &\xi^{\alpha\beta\gamma}\partial_\beta H_\gamma = \partial_t D^\alpha + J^\alpha \\ &\partial_\alpha D^\alpha = \rho \\ &\partial_\alpha B^\alpha = 0 \end{aligned} \tag{8.139}$$

Post's formalism preserves the affine nature of Maxwell' equations: their expression is independent of the coordinate system. The geometry only appears in the constitutive equations along with the material's properties

$$D^\alpha = \varepsilon^{\alpha\beta} E_\beta \qquad B^\alpha = \mu^{\alpha\beta} H_\beta \tag{8.140}$$

In a perfectly linear, isotropic media with permittivity ε and μ, these relation ships become:

$$\varepsilon^{\alpha\beta} = \varepsilon\sqrt{g}g^{\alpha\beta} \qquad \varepsilon^{\alpha\beta} = \varepsilon\sqrt{g}g^{\alpha\beta} \tag{8.141}$$

where $g^{\alpha\beta}$ are the contravariant components of the metric tensor.

$$\left(g^{\alpha\beta}\right) = \left(g_{\alpha\beta}\right)^{-1} \qquad g = (det)(g_{\alpha\beta}) \tag{8.142}$$

In an arbitrary curvilinear coordinates systeme x^α, if the surface separating two materials, denoted (1) and (2), coincides with a surface of coordinates $x^3 = constant$, for example, then the conditions of continuity are expressed quite simply

$$\text{tangential component continuity: } \begin{cases} a_1(1) = a_1(2) \\ a_2(1) = a_2(2) \end{cases} \tag{8.143}$$

$$\text{normal component continuity: } a^3(1) = a^3(2) \tag{8.144}$$

Assuming a time dependence of the form $\exp(-i\omega t)$, in a source free region if we substitute the constitutive equations for the material8.141 into Maxwell equations in the covariant form 8.139, setting

$$\mathscr{F}_\alpha = E_\alpha \qquad \mathscr{G}_\alpha = -iZH_\alpha \qquad \text{with } Z = \sqrt{\frac{\mu}{\varepsilon}} \tag{8.145}$$

we then obtain a set of equations relating the complex amplitudes of the field components where the $\mathscr{F}_\alpha$ and the $\mathscr{G}_\alpha$ play a fully symmetric role:

$$\begin{aligned} &\xi^{\alpha\beta\gamma}\partial_\beta \mathscr{F}_\gamma = k\sqrt{g}g^{\alpha\beta}\mathscr{G}_\alpha \\ &\xi^{\alpha\beta\gamma}\partial_\beta \mathscr{G}_\gamma = k\sqrt{g}g^{\alpha\beta}\mathscr{F}_\alpha \end{aligned} \tag{8.146}$$

where $k = \omega\sqrt{\mu\varepsilon}$

References:

8. 1. J. Chandezon, D. Maystre, and G. Raoult, “A new theoretical method for diffraction gratings and its numerical application J. Opt. (Paris) 11, 235–241 (1980).

8. 2. J. Chandezon, M. T. Dupuis, G. Cornet, and D. Maystre, “Multicoated gratings: a differential formalism applicable in the entire optical region,” J. Opt. Soc. Am. 72, 839–846 (1982).

8. 3. L. Li, J. Chandezon, G. Granet, and J. P. Plumey, “Rigorous and efficient grating-analysis method made easy for optical engineers,” Appl. Opt. 38, 304–313 (1999).

8. 4. E.J. Post, Formal structure of electromagnetics ,North Holland 1962.

8. 5. K. Edee, J.P.Plumey, J.Chandezon On the Rayleigh–Fourier method and the Chandezon method: Comparative study" Optics Communications , 286, pp 34-41 (2013).

8. 6. J.P. Plumey, G.Granet, and J.Chandezon,”Differential covariant formalism for solving Maxwell’s equations in curvilinear coordinates:oblique scattering from lossy periodic surfaces” IEEE Trans. Antennas Propag. 43, 835-842 (1995)

8. 7. L.Li,« Multilayer-coated diffraction gratings: differential method of Chandezon et al. revisited” , J.Opt.Soc.Am A11, 2816-2828 (1994).

8. 8. L. Li, “Justification of matrix truncation in the modal methods of diffraction gratings,” J. Opt. A, Pure Appl. Opt. 1, pp 531-536 (1999).

8. 9. "Numerical Techniques for microwave and millimeter-wave passive structures" Tatsuo Itoh editor.

8. 10. R. Redheffer, On the relation of transmission line theory to scattering and transfer, J. Mathematics and Physics 41, 1 – 41 (1962).

8. 11. G.Granet, J.-P.Plumey, and J.Chandezon,”Scattering by a periodically corrugated dielectric layer with non identical faces”, Pure Appl.Opt. 4, 1-5 (1995).

8. 12. L.Li,G.Granet,J-P.Plumey,and J.Chandezon,”some topics in extending the C method to multilayer-coated gratings of different profiles Pure Appl.Opt. 5,141-156 (1996).

8. 13. T.W. Preist,N.P.K.Cotter, and J.R.Sambles, J.Opt.Soc.Am.A, 12,pp 1740-1749 (1995).

8. 14. J.-P.Plumey, B.Guizal, and J.Chandezon,”Coordinate transformation method as applied to asymmetric gratings with vertical facets”, J.Opt.Soc.Am.A 14,610-617 (1997)

8. 15. T.W.Preist,J.B.Harris,N.P.Wanstall, and J.R.Sambles,"Optical response of blazed and overhanging gratings using Chandezon transformations", J.Mod.Opt.44,1073-1080 (1997).

8. 16. L. Li, "Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings," J. Opt. Soc. Am. A 13, pp 1024-1035 (1996)

8. 17. N.P.K. Cotter,T.W.Preist, and J.R.Sambles,"scattering matrix approach to multilayer diffraction," J.Opt. Soc.Am.A 12,pp 1097-1103 (1995)

8. 18. G. Granet, "Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution," J. Opt.Soc. Am. A 16, 2510–2516 (1999).

8. 19. G.Granet, J.Chandezon, J.P.Plumey, K.Raniriharinosy, "Reformulation of the coordinate transformation method through the concept of adaptive spatial resolution. Application to trapezoidal gratings," J. Opt. Soc. Am. A. 18, pp. 2102-2108 (2001).

8. 20. G. Granet and B. Guizal, "Analysis of strip gratings using a parametric modal method by Fourier expansions," Opt. Commun. 255, 1–11 (2005).

8. 21. Th. Weiss, G. Granet; N. Gippius, S. Tikhodeev, and H. Giessen, "Matched coordinates and adaptive spatial resolution in the Fourier modal method," Opt. Express 17, pp 8051-8061 (2009)

8. 22. L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," J. Opt.Soc. Am. A 13, 1870-1876 (1996)

8. 23. L. Li, "Oblique-coordinate-system-based Chandezon method for modeling one- dimensinally periodic, multilayer, inhomogeneous, anisotropic gratings," J. Opt. Soc. A 16, 2521-2531 (1999).

8. 24. R. Redheffer, On the relation of transmission line theory to scattering and transfer, J. Mathematics and Physics 41, 1 – 41 (1962).

8. 25. Numerical Techniques for microwave and millimeter-wave passive structures Tatsuo Itoh editor.

8. 26. R. Redheffer, Difference equations and functional equations in transmission-line theory, in E. F. Beckenbach (Ed.) Modern Mathematics for the Engineer, New York, McGraw-Hill, 1961.

Chapter 9:

Finite Difference Time Domain Method for Grating Structures

Fadi Issam Baida

and

Abderrahmane Belkhir

Table of Contents:

Chapter 9

Finite Difference Time Domain Method For Grating Structures

Fadi Issam Baida[1] and Abderrahmane Belkhir[2]

[1] *Institut FEMTO-ST, Département d'Optique P.M. Duffieux, UMR 6174 CNRS*
Université de Franche–Comté, 25030 Besançon Cedex, France
[2] *Université Mouloud Mammeri, Laboratoire de Physique et Chimie Quantique,*
Tizi-Ouzou, Algeria
fbaida@univ-fcomte.fr

The Finite Difference Time Domain method (FDTD), based on the Yee's scheme, is one of the most commonly used time methods for the modeling of electromagnetic waves propagation and diffraction. It was first introduced by Yee in 1966 [1] in the context of differential equations resolution and the first articles recommending its futur applications are published from 1975 [2, 3, 4]. Due to the simplicity of its implementation and the rapid growth of computing capacity, the FDTD is gaining users in all areas of electromagnetism applications. It allows a real-time monitoring of the electromagnetic wave evolution in any kind of environment (dielectric, metal, plasma...). Its theoretical formulation is very easy since it requires no matrix inversion and could take into account the more complex geometric shapes of objects in the studied system. In addition, using this time domain method, a wide spectral range characterization can be obtained from one temporal calculation via a simple Fourrier transform.

In this chapter, we present a brief review on the fundamentals of the FDTD method. We show how to adapt it to the calculation of the photonic band gap structures in the case of 2D periodic (invariant in the third direction) structures. The both in-plane, for the TE and TM polarizations, and off-plane propagations are considered. The last part of this chapter is devoted to FDTD general formulation, based on the Split Field Method technique, for the modeling of bi-periodic gratings that are finished according to the third direction.

9.1 Fundamentals of the FDTD method

9.1.1 The Yee's algorithm

The FDTD method is based on the numerical resolution of the Maxwell's equations using a centered finite difference schema to approximate the partial derivatives both in time and space.

Let us start from these equations expressed in their differential formulation:

$$\nabla \times \overrightarrow{E} = -\frac{\partial \overrightarrow{B}}{\partial t} \tag{9.1}$$

$$\nabla \times \overrightarrow{H} = \frac{\partial \overrightarrow{D}}{\partial t} \tag{9.2}$$

The electromagnetic properties of the medium are described through the so-called constitutive relationships:

$$\overrightarrow{D} = \varepsilon \overrightarrow{E} \tag{9.3}$$

$$\overrightarrow{B} = \mu \overrightarrow{H} \tag{9.4}$$

ε and μ are respectively the dielectric permittivity and magnetic permeability of the medium.

In a Cartesian coordinate system (O, x, y, z), the Maxwell's equations in the time domain are written as:

$$\frac{\partial H_x}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y}\right] \tag{9.5.a}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right] \tag{9.5.b}$$

$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left[\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}\right] \tag{9.5.c}$$

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}\right] \tag{9.5.d}$$

$$\frac{\partial E_y}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}\right] \tag{9.5.e}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right] \tag{9.5.f}$$

The numerical treatment of the partial differential equations 9.5 requires a space and time discretization. The calculation volume, shown in figure 9.1 is a rectangular parallelepiped divided into $(N_x \times N_y \times N_z)$ cells, each one with elementary volume $(\Delta x \times \Delta y \times \Delta z)$ where Δx, Δy and Δz are the spatial discretization steps according to the Ox, Oy and Oz directions respectively.

Each well defined node of the grid is associated with a triplet of integers (i, j, k) so that the coordinates $\left(x_i, y_j, z_k\right)$ of the node satisfy:

$$\begin{aligned} x_i &= i \cdot \Delta x \\ y_j &= j \cdot \Delta y \\ z_k &= k \cdot \Delta z \end{aligned}$$

The computational time is also discretized with a Δt time step. Each computing time t is associated with the integer n defining the number of temporal sampling:

$$t = n \cdot \Delta t$$

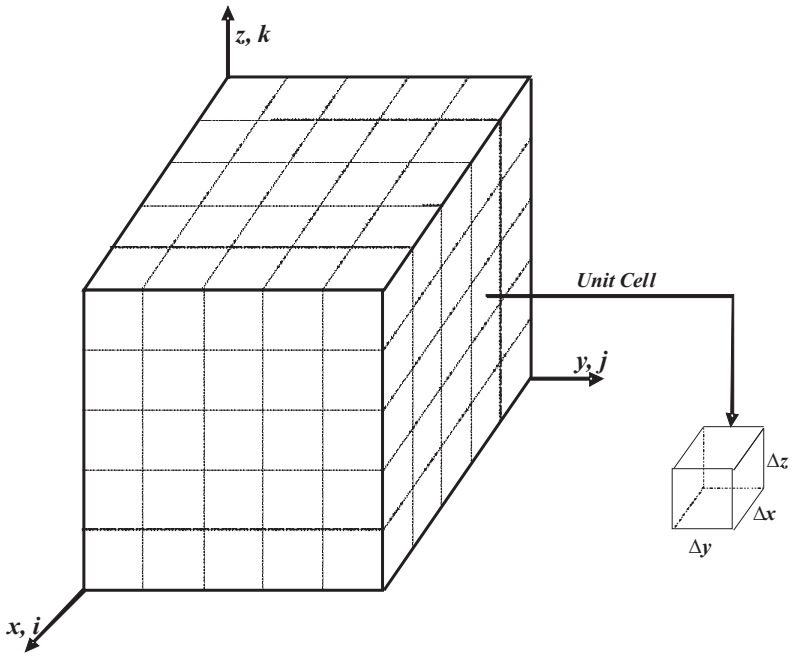


Figure 9.1: An exemple of the FDTD calculation volume.

Temporal and spatial derivatives of the field components (E_x, E_y, E_z, H_x, H_y, H_z) are approximated from their Taylor development to the first order. Thus, if U is one of these components, we will adopt the following notation:

$$U\left(x_i, y_j, z_k, t\right) = U^n_{i,j,k} \tag{9.6}$$

The temporal derivative of the U component at t time and $\left(x_i, y_j, z_k\right)$ node is approximated with finite centred difference as follows:

$$\left[\frac{\partial U}{\partial t}\right]_{i,j,k} = \frac{U^{n+\frac{1}{2}}_{i,j,k} - U^{n-\frac{1}{2}}_{i,j,k}}{\Delta t} + 0\left(\left[\Delta t\right]^2\right) \tag{9.7}$$

The spatial derivatives of the U component are approximated in the same manner:

$$\left[\frac{\partial U}{\partial x}\right]_{j,k,n} = \frac{U^n_{i+\frac{1}{2},j,k} - U^n_{i-\frac{1}{2},j,k}}{\Delta x} + 0\left(\left[\Delta x\right]^2\right) \tag{9.8.a}$$

$$\left[\frac{\partial U}{\partial y}\right]_{i,k,n} = \frac{U^n_{i,j+\frac{1}{2},k} - U^n_{i,j-\frac{1}{2},k}}{\Delta y} + 0\left(\left[\Delta y\right]^2\right) \tag{9.8.b}$$

$$\left[\frac{\partial U}{\partial z}\right]_{i,j,n} = \frac{U^n_{i,j,k+\frac{1}{2}} - U^n_{i,j,k-\frac{1}{2}}}{\Delta z} + 0\left(\left[\Delta z\right]^2\right) \tag{9.8.c}$$

As explicitly mentioned in equations 9.8, the use of centered difference scheme allows a precision of the second order even if a first order Taylor development is considered. This greatly enhances the numerical convergence of the FDTD algorithm.

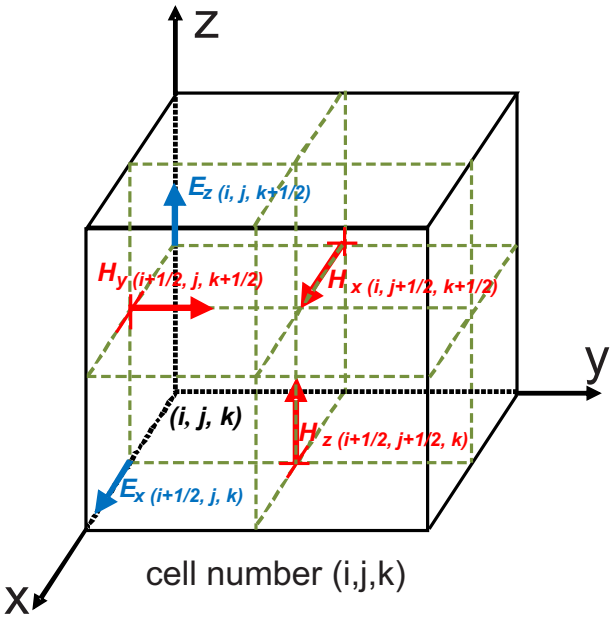


Figure 9.2: Spatial discretization : Yee's cell.

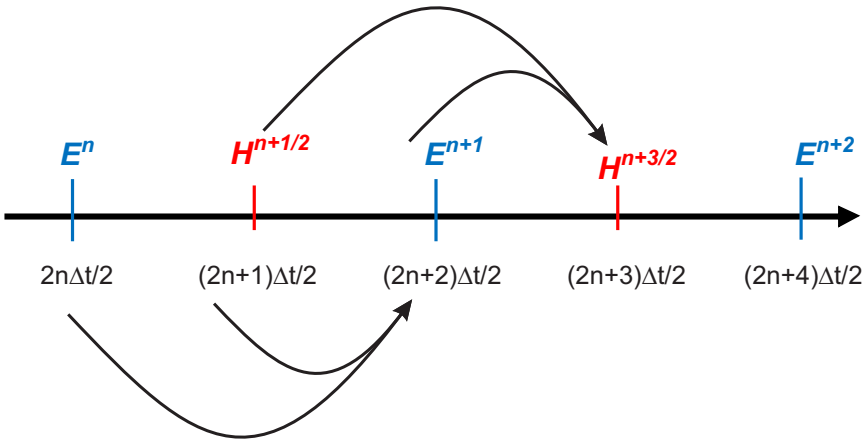


Figure 9.3: Temporal discretization into the Yee's scheme.

Yee's algorithm

The algorithm proposed by Kane Yee in 1966 [1] uses in a clever way this discretization for solving the system of equations (9.5). In the Yee's scheme, the electromagnetic field components are located at different points in a unit cell (Figure9.2). The electric field components are calculated along the edges of the cell while the perpendicular magnetic field components are calculated at the centers of the cell faces. Thus, each electric field component is surrounded by four magnetic field components and similarly for each magnetic field component.

The temporal increment into the Yee's scheme is done through a "leapfrog" discretization schema. The field components $\overrightarrow{H}$ (or $\overrightarrow{E}$) are calculated at times odd multiples of the half time step $\frac{\Delta t}{2}$, while the field components $\overrightarrow{E}$ (respectively $\overrightarrow{H}$) are updated at the times even multiples of $\frac{\Delta t}{2}$ as shown in figure 9.3. Such a discretization allows evaluating the time derivatives by keeping a centered difference schema as for spatial derivatives.

Consequently, replacing the partial derivatives in equations (9.5) by central difference (9.7-9.8), according to the Yee's scheme leads to the updated equations of electromagnetic components in the FDTD algorithm:

$$H_{x_{\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}}^{n+\frac{1}{2}} = H_{x_{\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}}^{n-\frac{1}{2}} - \frac{\Delta t}{\mu_0 \Delta}\left\{\left[E_{z_{\left(i,j+1,k+\frac{1}{2}\right)}}^{n} - E_{z_{\left(i,j,k+\frac{1}{2}\right)}}^{n}\right] + \left[E_{y_{\left(i,j+\frac{1}{2},k\right)}}^{n} - E_{y_{\left(i,j+\frac{1}{2},k+1\right)}}^{n}\right]\right\} \quad (9.9.a)$$

$$H_{y_{\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}}^{n+\frac{1}{2}} = H_{y_{\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}}^{n-\frac{1}{2}} - \frac{\Delta t}{\mu_0 \Delta}\left\{\left[E_{x_{\left(i+\frac{1}{2},j,k+1\right)}}^{n} - E_{x_{\left(i+\frac{1}{2},j,k\right)}}^{n}\right] + \left[E_{z_{\left(i,j,k+\frac{1}{2}\right)}}^{n} - E_{z_{\left(i+1,j,k+\frac{1}{2}\right)}}^{n}\right]\right\} \quad (9.9.b)$$

$$H_{z_{\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}}^{n+\frac{1}{2}} = H_{z_{\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}}^{n-\frac{1}{2}} - \frac{\Delta t}{\mu_0 \Delta}\left\{\left[E_{y_{\left(i+1,j+\frac{1}{2},k\right)}}^{n} - E_{y_{\left(i,j+\frac{1}{2},k\right)}}^{n}\right] + \left[E_{x_{\left(i+\frac{1}{2},j,k\right)}}^{n} - E_{x_{\left(i+\frac{1}{2},j+1,k\right)}}^{n}\right]\right\} \quad (9.9.c)$$

$$E_{x_{\left(i+\frac{1}{2},j,k\right)}}^{n+1} = E_{x_{\left(i+\frac{1}{2},j,k\right)}}^{n} + \frac{\Delta t}{\varepsilon \Delta}\left\{\left[H_{z_{\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}}^{n} - H_{z_{\left(i+\frac{1}{2},j-\frac{1}{2},k\right)}}^{n}\right] + \left[H_{y_{\left(i+\frac{1}{2},j,k-\frac{1}{2}\right)}}^{n} - H_{y_{\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}}^{n}\right]\right\} \quad (9.9.d)$$

$$E_{y_{\left(i,j+\frac{1}{2},k\right)}}^{n+1} = E_{y_{\left(i,j+\frac{1}{2},k\right)}}^{n} + \frac{\Delta t}{\varepsilon \Delta}\left\{\left[H_{x_{\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}}^{n} - H_{x_{\left(i,j+\frac{1}{2},k-\frac{1}{2}\right)}}^{n}\right] + \left[H_{z_{\left(i-\frac{1}{2},j+\frac{1}{2},k\right)}}^{n} - H_{z_{\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}}^{n}\right]\right\} \quad (9.9.e)$$

$$E_{z_{\left(i,j,k+\frac{1}{2}\right)}}^{n+1} = E_{z_{\left(i,j,k+\frac{1}{2}\right)}}^{n} + \frac{\Delta t}{\varepsilon \Delta}\left\{\left[H_{y_{\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}}^{n} - H_{y_{\left(i-\frac{1}{2},j,k+\frac{1}{2}\right)}}^{n}\right] + \left[H_{x_{\left(i,j-\frac{1}{2},k+\frac{1}{2}\right)}}^{n} - H_{x_{\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}}^{n}\right]\right\} \quad (9.9.f)$$

Let us note that this last equation system can be simplified significantly in case of 2D structures (see section 2 of this chapter).

For the modeling of structures with a symmetry of revolution, a basis change from Cartesian to cylindrical coordinates is strongly recommended to accurately describe the fine details of the samples and to make more flexible the FDTD calculation codes. In these so-called BOR-FDTD (Body of Revolution FDTD) codes, the symmetry of revolution is exploited to express

the azimuthal dependence (ϕ) of the electromagnetic fields as Fourrier series. BOR-FDTD algorithm can, in this case, compute solutions for all Fourier modes through one simulation per mode. This code is commonly called 2.5D since the azimuthal field variation is analytically accounted for. Thus, there is no griding in the ϕ-direction. This implies that the BOR-FDTD algorithm is two-dimensional in terms of computer ressource usage even 3D structures are modeled.

9.1.2 Spatiotemporal criteria of convergence

As all explicit schemes, Yee's algorithm is subjected to a stability condition setting the time step from the space discretization. Arbitrary values of spatiotemporal discretization can lead to infinite solutions of the electromagnetic field. Stability problems in explicit numerical methods have been analyzed in detail by Courant, Friedrichs and Levy [5] and Von Neumann, from a mathematically rigorous approach. This analysis shows that the explicit schemes are stable under a condition called CFL (for Current, Friedrich and Levy) and applied to the FDTD method in the case of a regular mesh [6]:

$$\Delta t \leq \left[v_{max} \cdot \sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} + \frac{1}{\Delta z^2}} \right]^{-1} \tag{9.10}$$

where v_{max} is the maximum velocity of light propagation in the studied system, generally the velocity of light in vacuum.

In case of uniforme mesh ($\triangle x = \triangle y = \triangle z = \triangle$), the CFL criterion becomes:

$$\Delta t \leq \frac{1}{v_{max}} \cdot \frac{\Delta}{\sqrt{3}} \qquad \text{in } 3D \tag{9.11}$$

$$\Delta t \leq \frac{1}{v_{max}} \cdot \frac{\Delta}{\sqrt{2}} \qquad \text{in } 2D \tag{9.12}$$

However, it is possible to overcome the restrictive assumption of regular mesh that achieves the above result with the following generalized criterion:

$$\Delta t \leq \left[v_{max} \cdot \sqrt{\frac{1}{\Delta x_{min}^2} + \frac{1}{\Delta y_{min}^2} + \frac{1}{\Delta z_{min}^2}} \right]^{-1} \tag{9.13}$$

where Δx_{min}, Δy_{min} et Δz_{min} are the smallest step in the three directions x, y and z respectively.

In addition to the numerical instability problem, the transition from continuous forms of Maxwell's equations to the discrete numerical approximations can cause a parasitic effect called "numerical dispersion". This is due to the fact that numerical signals are propagated over time in the FDTD grid, with a phase velocity less than the actual velocity. This dispersion varies with frequency, propagation direction in the grid and the spatial discretization [6]. Numerical dispersion errors increase with the signal frequency and size of the computational domain, thus

making the simulation results less reliable. They may appear in various forms: phase error, signal distortion, loss of amplitude, pulse broadening ...

The solution to this problem requires a very fine mesh in the FDTD grid, so that the maximum discretization is of the order $\lambda_{min}/20$ [6], λ_{min}being the minimum wavelength of propagating waves in the FDTD grid.

9.1.3 Absorbing boundary conditions - Perfectly Matched Layers

Such conditions allow us to describe open systems where emitted or reflected waves propagate to infinity. Indeed, the limited memory space of computers requires users to truncate their FDTD computational domain. At the limits of this truncated domain, components of the electromagnetic field can not be calculated by the discretized equations (9.9). Therefore special treatment at the borders is needed to avoid the incident electromagnetic wave on these "edges" does reflect back and contaminate the actual physical signal. One of the most widely used technique is that proposed by Berenger [7] called Perfectly Matched Layer (PML). This technique consists of adding around the studied domain not necessarily physical layer causing no reflection and almost totally absorbing all the propagating electromagnetic field. Its use is based on the condition of impedance matching of two waves at the interface between two media with the same index but which one is absorbing (with nonzero electrical conductivity σ and magnetic equivalent conductivity $\sigma^\star$ as shown in figure 9.4).

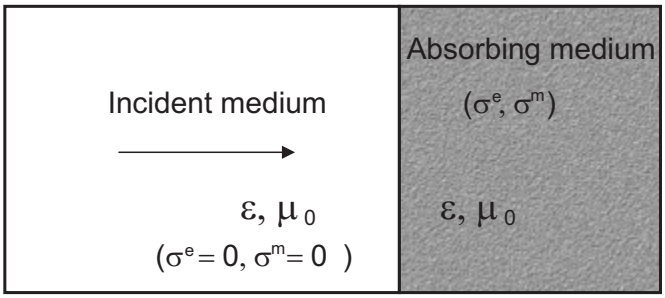


Figure 9.4: Impedance matching principle.

This condition is expressed as:

$$\frac{\sigma}{\varepsilon} = \frac{\sigma^\star}{\mu_0} \tag{9.14}$$

Thus, a magnetic conductivity is needed to fulfill this impedance matching condition. In addition, absorption is needed only for components of the fields that propagates perpendicularly to the interface (the FDTD window border) and not in the parallel direction. Bérenger solved this problem by proposing an artificially biaxial absorbing medium. The absorption is not zero in the direction normal to the interface between the two media and is zero along the axis parallel to the interface. In the PML medium, the incident plane wave is split into two fictitious waves (see figure 9.5):

1) A wave propagating at normal incidence and satisfying the equation 9.14. This wave is attenuated and absorbed by the PML medium and undergoes only very low reflectivity to the incident medium.

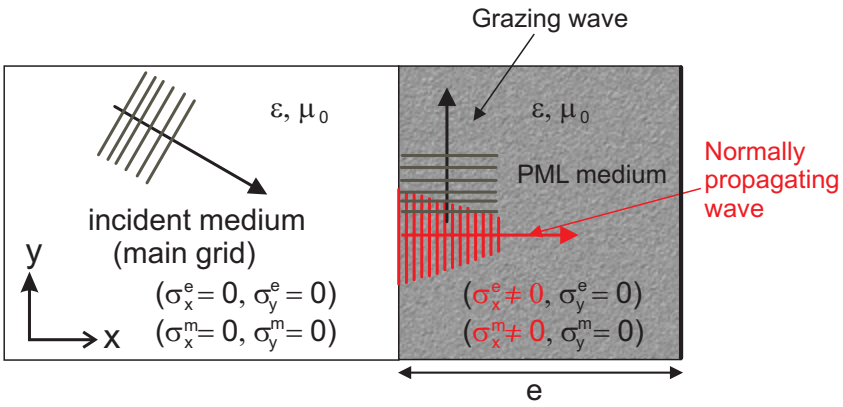


Figure 9.5: Schematic of the PML principle.

2) A second grazing incidence wave that shows no absorption in the PML medium. This wave, propagating parallel to the interface between two media undergoes no reflection and sees a medium identical to that of the main grid window.

Abrupt changes in conductivities at this interface degrade the performances of absorption. This effect is, however, reduced by imposing a progressive variation of the absorption according to a polynomial law given by [7]:

$$\sigma = \sigma_{max} \left(\frac{x_{pml}}{e} \right)^m \tag{9.15}$$

where σ_{max} is the maximum value of the conductivity, x_{pml} represents the depth in the PML region measured from the interface, e denotes the thickness of the PML layer and m is the polynomial order generally fixed to 2.

Let us note that in the case of gratings such conditions are not necessary according to the periodicity directions. The absorbing boundaries conditions are hereby replaced by Floquet-Bloch periodic conditions in order to describe periodic structures (see section 2 of this chapter). Nevertheless, for a 2D periodic structure, PML are needed in the third direction where the structure is usually finite.

9.1.4 Dispersive media

The dispersive media, such as metals in the optical range, are characterized by a complex permittivity frequency dependent $\varepsilon(\omega) = \varepsilon'(\omega) + i\varepsilon''(\omega)$. As the FDTD method is temporal, in such environments the direct implementation of the above equations, in which appear explicitly permittivity and hence the frequency, is impossible. The solution for this problem is to calculate the displacement vector $\overrightarrow{D}$ components in the classical Yee's scheme and then back to electrical field components using the constitutive equation of the medium established in the frequency domain $\overrightarrow{D}(\omega) = \varepsilon(\omega)\overrightarrow{E}(\omega)$. The temporal nature of the FDTD needs a temporal constitutive equation written as a convolution product $\overrightarrow{D}(t) = \varepsilon(t) \otimes \overrightarrow{E}(t)$. It is a non local relationship whose resolution requires the knowledge of the electric field at all previous times. Numerically, this leads to a storage of a very large amount of data and therefore requires to have a very large memory space. This issue can be bypassed using analytical models describing the dielectric function $\varepsilon(\omega)$ of these metals. The choice of adapted analytical model depends on the type of metal as well as the spectral range of study.

9.1.4.1 Drude Model

The Drude model of free electrons [8, 9] for the dielectric function which, although based on a purely classical approach, can well account for intraband transitions. In this model, firstly proposed in 1908 by P. Drude, a gas of free electrons moving in a immobile metal ions lattice. Thus, the electron-electron interactions and electron-ions are not taken into account and the movement of all the electron cloud is thus the average of the movements of individual electrons. The relative permittivity given by this model is:

$$\varepsilon_D = \varepsilon_\infty - \frac{\omega_D^2}{\omega^2 + i\omega\gamma_D} \tag{9.16}$$

where ω_D is the "plasma frequency" of the metal and ε_∞ its relative permittivity at infinite frequencies. γ_D represents a damping term that is inversely proportional to the relaxation time.

FDTD implementation of the Drude model

The principle consists in replacing the electric field vector $\overrightarrow{E}$ by $\overrightarrow{D}/\varepsilon$ in Maxwell's equations in order to eliminate ε term. In dispersive media, equations (9.9.d, 9.9.e et 9.9.f) are replaced by:

$$D_{x\left(i+\frac{1}{2},j,k\right)}^{n+1} = D_{x\left(i+\frac{1}{2},j,k\right)}^{n} + \frac{\Delta t}{\Delta}\left\{\left[H_{z\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}^{n} - H_{z\left(i+\frac{1}{2},j-\frac{1}{2},k\right)}^{n}\right] + \left[H_{y\left(i+\frac{1}{2},j,k-\frac{1}{2}\right)}^{n} - H_{y\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}^{n}\right]\right\} \tag{9.17}$$

$$D_{y\left(i,j+\frac{1}{2},k\right)}^{n+1} = E_{y\left(i,j+\frac{1}{2},k\right)}^{n} + \frac{\Delta t}{\Delta}\left\{\left[H_{x\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}^{n} - H_{x\left(i,j+\frac{1}{2},k-\frac{1}{2}\right)}^{n}\right] + \left[H_{z\left(i-\frac{1}{2},j+\frac{1}{2},k\right)}^{n} - H_{z\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}^{n}\right]\right\} \tag{9.18}$$

$$D_{z\left(i,j,k+\frac{1}{2}\right)}^{n+1} = D_{z\left(i,j,k+\frac{1}{2}\right)}^{n} + \frac{\Delta t}{\Delta}\left\{\left[H_{y\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}^{n} - H_{y\left(i-\frac{1}{2},j,k+\frac{1}{2}\right)}^{n}\right] + \left[H_{x\left(i,j-\frac{1}{2},k+\frac{1}{2}\right)}^{n} - H_{x\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}^{n}\right]\right\} \tag{9.19}$$

Once the components of the displacement vector $\overrightarrow{D}$ are updated from the previous equations, we proceed to the determination of the $\overrightarrow{E}$ components using the relation $\overrightarrow{D} = \varepsilon(\omega)\overrightarrow{E}$. Replacing $\varepsilon(\omega)$ by its expression given by the Drude model, we get to:

$$(\omega^2 + i\omega\gamma_D)\overrightarrow{D} = \varepsilon_0\varepsilon_\infty(\omega^2 + i\omega\gamma_D)\overrightarrow{E} - \varepsilon_0\omega_D^2\overrightarrow{E} \tag{9.20}$$

Assuming time dependance of electromagnetic field in $e^{-i\omega t}$, a simple Fourier transform ($\omega \rightarrow t$) of this last equation leads to:

$$\frac{d^2\overrightarrow{D}}{dt^2} + \gamma_D \frac{d\overrightarrow{D}}{dt} = \varepsilon_0(\varepsilon_\infty \frac{d^2\overrightarrow{E}}{dt^2} + \varepsilon_\infty \gamma_D \frac{d\overrightarrow{E}}{dt} + \omega_D^2 \overrightarrow{E})$$

The partial derivatives of this equations are then replaced by their expressions through the centered finite difference schema. The electric field updated equation in the dispersive media is then obtained:

$$\xi \overrightarrow{E}^{n+1} = -\chi \overrightarrow{E}^{n-1} + 4\varepsilon_\infty \varepsilon_0 \overrightarrow{E}^n + \overrightarrow{D}^{n+1}[\gamma_D \Delta t + 2] - 4\overrightarrow{D}^n + [-\gamma_D \Delta t + 2]\overrightarrow{D}^{n-1} \tag{9.21}$$

with $\xi = \varepsilon_0[\omega_D^2 \Delta t^2 + \varepsilon_\infty \gamma_D \Delta t + 2\varepsilon_\infty]$ and $\chi = \varepsilon_0[\omega_D^2 \Delta t^2 - \gamma_D \varepsilon_\infty \Delta t + 2\varepsilon_\infty]$. Due to the dispersion, an additional step of calculation is necessary. It consists of determining the displacement field components for all nodes representing the dispersive media. In addition and as can be seen in equation (9.21), we need to store the $\overrightarrow{E}$ and $\overrightarrow{D}$ components on two time steps, which has the effect of increasing the memory space to be allocated and the computation time.

9.1.4.2 Drude-Lorentz Model

In addition to the conduction electrons, the Drude-Lorentz model takes into account the bound electrons. The interband transition of electrons from filled bands to the conduction band can significantly influence the optical response. In alkali metals, these transitions occur at high frequencies and provide only small corrections to the dielectric function in the optical domain. These metals are well described by the Drude model. On the other side, in noble metals a correction must be made to the dielectric function. It is due to transitions between the bands d and the conduction band s-p. The contribution of bound electrons to the dielectric function can be described by the Lorentz model. To the above Drude dielectric function, a Lorentzian term is added:

$$\varepsilon_{DL}(\omega) = \varepsilon_D(\omega) + \varepsilon_L(\omega)$$

Estimating $\varepsilon_L(\omega)$, the bound electrons are described by forced and damped harmonic oscillators. Vial *et al*. [10] suggested a single oscillator leading to a single Lorentzian additional term to well describe the permittivity of gold in the optical range compared with the classical Drude model. In this case, the relative dielectric function is:

$$\varepsilon_{DL}(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\omega\gamma} - \frac{\Delta\varepsilon \cdot \Omega_L^2}{(\omega^2 - \Omega_L^2) + i\Gamma_L\omega} \tag{9.22}$$

where Γ_L et Ω_L stand for the spectral width and the strength of the Lorentz oscillator respectively. $\Delta\varepsilon$ is a weighting factor.

The FDTD implementation of this model can be done with the Auxilliary Differential Equations (ADE) method previously described above in the case of the Drude model or the so-called Recursive Convolution (RC) method [10]. Because of the additional Lorentzian term, its use requires the introduction of additional intermediate electromagnetic components in the algorithm. Thus, a larger memory space is required compared to the case of the Drude model. In general, many involving multiple oscillators Lorentz terms are needed to accurately model the permittivity of noble metals in the optical range.

9.1.4.3 Drude critical points model

The optical properties of some metals, particularly gold, are more difficult to be analytically described in the visible/near-UV region. This comes from much more important role, in the case of gold, played by interband transitions in this region. Some attempts to add Lorentz oscillators to the classical Drude term to account for these transitions rapidly face limitations [11]. In fact, besides the huge simulation time, increasing the number of parameters (mainly non-physical and not well defined) would not provide more insight than quality fit (itself non-physical) with a polynomial high degree or a simple numerical interpolation of the experimental data.

In order to achieve a reasonable representation of the dielectric function, Etchegoin *et al*. [12] took inspiration from the parametric critical points model developed for semiconductors [13]. This model is very suitable for the description of optical properties of metals (such as gold) for which the band structure is quite complex. In this approach, the frequency dependence of the optical properties of gold in the visible/near-UV may be well described by an analytical formula with three main contributions that can be expressed as follows:

$$\varepsilon_{D2CP}(\omega) = \varepsilon_\infty - \frac{\omega_D^2}{\omega^2 + i\omega\gamma_D} + \sum_{p=1}^{p=2} G_p(\omega) \tag{9.23}$$

with

$$G_p(\omega) = A_p \Omega_p \left(\frac{e^{i\phi_p}}{\Omega_p - \omega - i\Gamma_p} + \frac{e^{-i\phi_p}}{\Omega_p + \omega + i\Gamma_p} \right) \tag{9.24}$$

The two first terms of equation (9.23) represents the standard contribution of the classical Drude Model. The sum in equation (9.23) is the contribution of the inter-band transitions with the amplitude A_p, gap energy Ω_p, phase ϕ_p and broadening Γ_p.

In a comparative study of this Drude critical points (CP) model with the so-called L4 model which consists of four Lorentzian terms [14], Vial *et al*. [15] show the possibility to increase the accuracy of gold and silver permittivity description by using the CP model with fewer parameters to determine and less memory use within the FDTD method.

Implementation of the CP model in FDTD using ADE technique

As in the previous case of the Drude model, the technique is to calculate the displacement vector components by the FDTD equations (9.17,9.18 and 9.19) and determine electrical components using the following relationship:

$$\overrightarrow{D} = \varepsilon_0 \varepsilon_{DCP} \overrightarrow{E} \tag{9.25}$$

In the case of the CP model, $\overrightarrow{D}$ can be written as the sum of the electric displacement vectors corresponding to each of the contributions in the dielectric function expression:

$$\overrightarrow{D} = \overrightarrow{D}_D + \sum_{p=1}^{2} \overrightarrow{D}_{Cp} \tag{9.26}$$

with

$$\overrightarrow{D}_D = \varepsilon_0[\varepsilon_\infty - \frac{\omega_P^2}{\omega^2 + \mathrm{i}\gamma\omega}]\overrightarrow{E} \tag{9.27.a}$$

$$\overrightarrow{D}_{Cp} = \varepsilon_0[A_p\Omega_p(\frac{\mathrm{e}^{\mathrm{i}\phi_p}}{\Omega_p - \omega - \mathrm{i}\Gamma_p} + \frac{\mathrm{e}^{-\mathrm{i}\phi_p}}{\Omega_p + \omega + \mathrm{i}\Gamma_p})]\overrightarrow{E} \tag{9.27.b}$$

As before the temporal evolution of the fields in $e^{-i\omega t}$ is considered. By inverse Fourier transform, we obtain:

$$(\frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t})\overrightarrow{D_D} = \varepsilon_0\varepsilon_\infty(\frac{\partial^2}{\partial t^2} + \gamma\frac{\partial}{\partial t} + \frac{\omega_P^2}{\varepsilon_\infty})\overrightarrow{E} \tag{9.28.a}$$

$$(\Omega_p^2 + \Gamma_p^2 + \frac{\partial^2}{\partial t^2} + 2\Gamma_p\frac{\partial}{\partial t})\overrightarrow{D_{Cp}} = 2\varepsilon_0 A_p\Omega_p(\sqrt{\Gamma_p^2 + \Omega_p^2}\sin(\theta_p - \phi_p) - \sin\phi_p\frac{\partial}{\partial t})\overrightarrow{E} \tag{9.28.b}$$

where: $\theta_p = \arctan(\frac{\Omega_p}{\Gamma_p})$

By centered difference discretization of the equation system (9.28) and taking into account the split equation of the displacement vector (9.26), we reach the updated equations system for the electric field vector at each point (i, j, k) of the calculation window:

$$\begin{aligned}\overrightarrow{E}^{n+1} = \frac{1}{\frac{\chi_D}{\alpha_D} + \sum_{p=1}^{p=2}(\frac{\chi_p}{\alpha_p})}\Bigg[&\overrightarrow{D}^{n+1} + \frac{\beta_D}{\alpha_D}\overrightarrow{D}_D^{n-1} + \frac{4}{\alpha_D}\overrightarrow{D}_D^n - \frac{\delta_D}{\alpha_D}\overrightarrow{E}^{n-1} - \frac{4\varepsilon_0\varepsilon_\infty}{\alpha_D}\overrightarrow{E}^n \\ &+ \sum_{p=1}^{p=2}(\frac{\beta_p}{\alpha_p}\overrightarrow{D}_{Cp}^{n-1} - \frac{4}{\alpha_p}\overrightarrow{D}_{Cp}^n) + \sum_{p=1}^{p=2}(\frac{\delta_p}{\alpha_p})\overrightarrow{E}^{n-1}\Bigg]\end{aligned} \tag{9.29.a}$$

$$\overrightarrow{D}_D^{n+1} = \frac{1}{\alpha_D}\left[-\beta_D\overrightarrow{D}_D^{n-1} - 4\overrightarrow{D}_D^n + \chi_D\overrightarrow{E}^{n+1} + \delta_D\overrightarrow{E}^{n-1} + 4\varepsilon_0\varepsilon_\infty\overrightarrow{E}^n\right] \tag{9.29.b}$$

$$\overrightarrow{D}_{Cp}^{n+1} = \frac{1}{\alpha_p}[-\beta_p\overrightarrow{D}_{Cp}^{n-1} + 4\overrightarrow{D}_{Cp}^n + \chi_p\overrightarrow{E}^{n+1} + \delta_p\overrightarrow{E}^{n-1}] \tag{9.29.c}$$

with:

$$\alpha_D = -2 - \gamma\Delta t \tag{9.30a}$$

$$\beta_D = -2 + \gamma\Delta t \tag{9.30b}$$

$$\chi_D = \varepsilon_0\,\varepsilon_\infty[-2 - \gamma\Delta t - (\omega_p\Delta t)^2/\varepsilon_\infty] \tag{9.30c}$$

$$\delta_D = \varepsilon_0\,\varepsilon_\infty[-2 + \gamma\Delta t - (\omega_p\Delta t)^2/\varepsilon_\infty] \tag{9.30d}$$

$$\alpha_p = [\Omega_p^2 + \Gamma_p^2]\Delta t^2 + 2\Gamma_p\Delta t + 2 \tag{9.30e}$$

$$\beta_p = [\Omega_p^2 + \Gamma_p^2]\Delta t^2 - 2\Gamma_p\Delta t + 2 \tag{9.30f}$$

$$\chi_p = 2A_p\Omega_p\varepsilon_0[\Delta t^2\sqrt{\Omega_p^2 + \Gamma_p^2}\sin(\theta_p - \phi_p) - \Delta t\sin\phi_p] \tag{9.30g}$$

$$\delta_p = 2A_p\Omega_p\varepsilon_0[\Delta t^2\sqrt{\Omega_p^2 + \Gamma_p^2}\sin(\theta_p - \phi_p) + \Delta t\sin\phi_p] \tag{9.30h}$$

Let us mention that the displacement vector split into three contributions avoids doing appear derivatives of order higher than 2 in the equations system (9.28). As seen on the equations

system (9.29), taking into consideration the two critical points in the FDTD algorithm does not need to store $\overrightarrow{E}$ and $\overrightarrow{D}$ components over more than two time steps. However, against the Drude model implementation, additional calculation stages appear in order to determine the two parts of the displacement vector corresponding to the two critical contributions.

9.2 Band gap calculation for 2D periodic structures

In this section, we describe how to adapt the FDTD calculation for photonic bandgap structures (PBG) of periodic arrays. The biperiodic structures case is there considered. These 2D structures are photonic crystals (PhC) whose permittivity is periodic in two dimensions (x and y for example) and remains invariant according to the third one (z). They mainly include three main families that are square, triangular and hexagonal lattices. For this type of structures, we can distinguish two kinds of propagation, in the plane (in-plane, $k_z = 0$) and out of plane (off-plane, nonzero k_z). The system of equations (9.5) becomes easier depending on the type of propagation. To illustrate this, let us assume in what follows that the PhC is periodic along the x and y directions and infinite along z direction.

9.2.1 In-plane propagation: TE and TM polarizations

In that case the propagation is done in the plane and the field variation vanishes along the third direction. The system of equations (9.5) is simplified and divided into two independent subsystems giving rise to two polarizations: transverse electric (TE) and transverse magnetic (TM):

TE Polarization

$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}\right) \tag{9.31a}$$

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\frac{\partial H_z}{\partial y} \tag{9.31b}$$

$$\frac{\partial E_y}{\partial t} = -\frac{1}{\varepsilon}\frac{\partial H_z}{\partial x} \tag{9.31c}$$

TM Polarization

$$\frac{\partial H_x}{\partial t} = -\frac{1}{\mu}\frac{\partial E_z}{\partial y} \tag{9.32a}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\frac{\partial E_z}{\partial x} \tag{9.32b}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right) \tag{9.32c}$$

In case of TE polarization, the electrical components are transverse. They are in the plane of periodicity of the PhC. On the other hand, for the TM polarization, the electric field is perpendicular to the directions of periodicity and the magnetic components are transverse.

Let us note that the two polarizations can be studied by the same system of equations (9.5) without separating it into two sub-systems. But to simplify the calculation codes and gain memory space, it is recommended to study these two polarizations separately.

9.2.2 Off-plane propagation

Off-plane propagation is characterized by a nonzero propagation constant k_z according to z direction. Diagram dispersion is generally determined for a fixed value of k_z. Thus, the z-derivatives in Maxwell equations become analytical while the electric and magnetic field vectors can be written as follows:

$$\overrightarrow{E}(x,y,z,t) = \overrightarrow{E_0}(x,y,t)\exp(ik_z z) \tag{9.33a}$$
$$\overrightarrow{H}(x,y,z,t) = \overrightarrow{H_0}(x,y,t)\exp(ik_z z) \tag{9.33b}$$

The Maxwell's system of equations (9.5) becomes:

$$\frac{\partial H_x}{\partial t} = \frac{1}{\mu}(ik_z E_y - \frac{\partial E_z}{\partial y}) \tag{9.34a}$$
$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}(\frac{\partial E_z}{\partial x} - ik_z E_x) \tag{9.34b}$$
$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}) \tag{9.34c}$$
$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}(\frac{\partial H_z}{\partial y} - ik_z H_y) \tag{9.34d}$$
$$\frac{\partial E_y}{\partial t} = \frac{1}{\varepsilon}(ik_z H_x - \frac{\partial H_z}{\partial x}) \tag{9.34e}$$
$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}) \tag{9.34f}$$

In this case, it is no longer possible to separate the system into two subsystems as before. The TE and TM cases are therefore mixed together and can not be treated separately. However, we can note that the calculation code is simplified since the z derivatives are analytically evaluated so there is no discretization along the z direction. A 2D algorithm is still needed.

9.2.3 Periodic boundary conditions

As the CPU time and space memory is limited, the FDTD calculation window must also be finite. Because of symmetry, only one unit cell is considered. To reproduce the crystal at the truncated domain boundaries, the Floquet-Bloch periodicity conditions [9] are applied to the electric and magnetic components. Despite the fact that these periodicity conditions are general and can be applied to any periodic structure, their expressions depend on the Bravais lattice. Consequently, we will consider the two most used Bravais lattices i.e. the rectangular and the triangular ones.

Rectangular cell

Let us consider a PhC made of cylinders (refractive index n_1) immersed in a medium of refractive index n_2. a and b are the lattice constants in the x and y directions respectively (see figure 9.6). The FDTD window calculation is shown in figure 9.6-b.

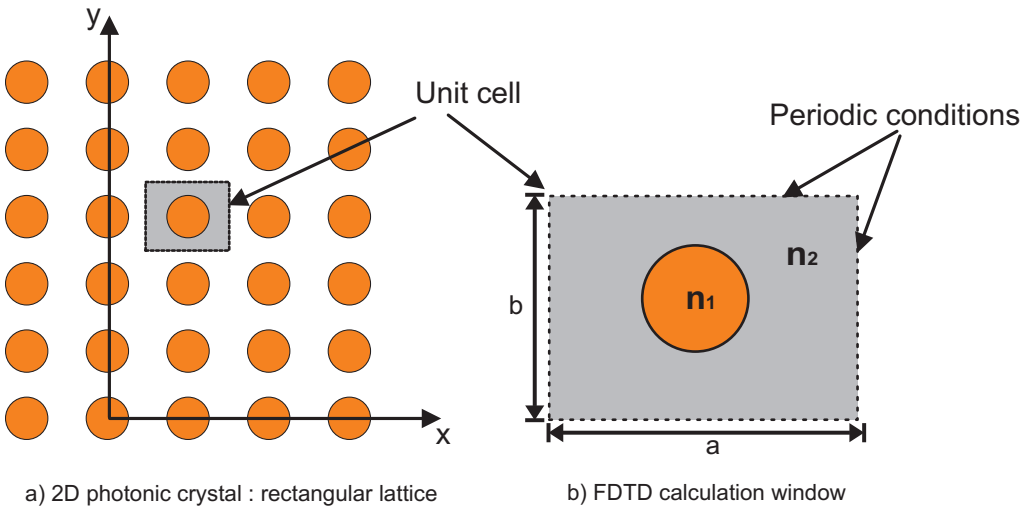


Figure 9.6: Rectangular structure and FDTD window calculation

The Floquet-Bloch conditions are applied to the electric and magnetic components as follows:

$$\overrightarrow{E}(x=0,y,t) = \overrightarrow{E}(x=a,y,t)\exp(-ik_x\cdot a) \quad (9.35a)$$
$$\overrightarrow{E}(x,y=0,t) = \overrightarrow{E}(x,y=b,t)\exp(-ik_y\cdot b) \quad (9.35b)$$
$$\overrightarrow{H}(x=a,y,t) = \overrightarrow{H}(x=0,y,t)\exp(ik_x\cdot a) \quad (9.35c)$$
$$\overrightarrow{H}(x,y=b,t) = \overrightarrow{H}(x,y=0,t)\exp(ik_y\cdot b) \quad (9.35d)$$

Triangular cell

Similarly to the rectangular cell, the calculation FDTD window is limited to a single unit cell. To model the triangular photonic structure (see figure 9.7-a), three choices of the FDTD window are possible. The first one is to take a non-orthogonal unit cell (cell 1 in figure 9.7-a) and implement the periodic boundary conditions in a Non orthogonal-FDTD algorithm [16, 17] for which the classical FDTD developed in an orthogonal coordinate system is not suitable. To bypass this constraint and remaining in the conventional FDTD with orthogonal coordinates, the second rectangular cell (celle 2 in figure 9.7-a) can be used to derive the periodic conditions. Nevertheless, this cell contains two patterns. This means that the rectangular periodic conditions lead to a less-description of all the possible solutions. Consequently, an aliasing effect will appear in the dispersion diagram.

In order to get gain in computational time and space and prevent this band folding while remaining with the orthogonal FDTD algorithm, a rectangular cell can be defined with only one pattern (cell 3 in figure 9.7-a). Within this FDTD calculation cell (9.7-b), the periodic conditions above are therefore replaced by:

-along the x direction :

$$\overrightarrow{E}(x=0,y,z,t) = \overrightarrow{E}(x=a,y,z,t)\exp(-ik_x\cdot a) \quad (9.36a)$$
$$\overrightarrow{H}(x=a,y,z,t) = \overrightarrow{H}(x=0,y,z,t)\exp(ik_x\cdot a) \quad (9.36b)$$

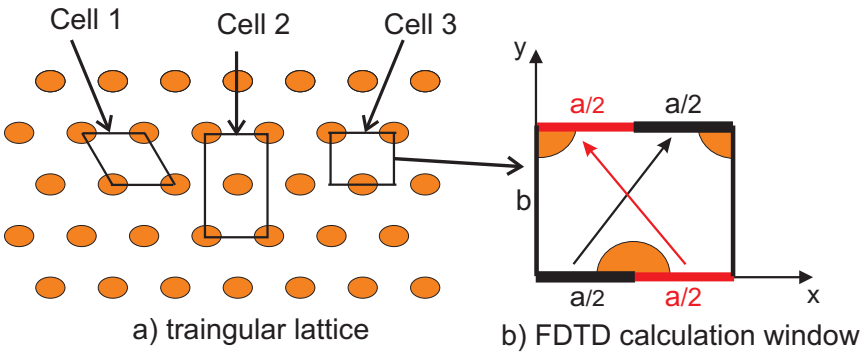


Figure 9.7: Triangular structure and FDTD calculation window

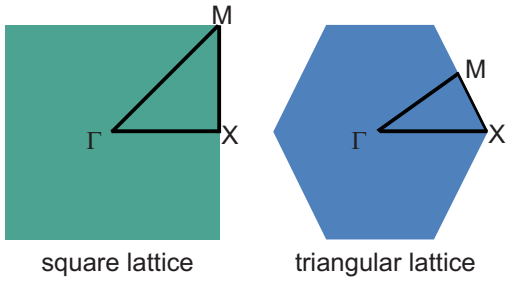


Figure 9.8: Brillouin zone

- along the y direction with $x \geq 0$ and $x \leq a/2$

$$\vec{E}(x, y=0, z, t) = \vec{E}(x+\frac{a}{2}, y=b, z, t)\exp(i(-k_y \cdot b - k_x \cdot a/2)) \quad (9.37a)$$

$$\vec{H}(x, y=b, z, t) = \vec{H}(x+\frac{a}{2}, y=0, z, t)\exp(i(k_y \cdot b - k_x \cdot a/2)) \quad (9.37b)$$

- along the y direction with $x > a/2$ and $x \leq a$

$$\vec{E}(x, y=0, z, t) = \vec{E}(x-\frac{a}{2}, y=b, z, t)\exp(i(-k_y \cdot b + k_x \cdot a/2)) \quad (9.38a)$$

$$\vec{H}(x, y=b, z, t) = \vec{H}(x-\frac{a}{2}, y=0, z, t)\exp(i(k_y \cdot b + k_x \cdot a/2)) \quad (9.38b)$$

By the way, the dispersion diagram of a triangular or honeycomb Bravais lattices can be calculated without modifying the orthogonal Cartesian Yee schema.

9.2.4 Some examples of band gap calculation

To obtain a photonic band diagram, several FDTD calculations are necessary done by varying the $\vec{k}$ wavevector that must scan the irreducible Brillouin zone (figure 9.8). ΓX, XM and $M\Gamma$ highest symmetry directions are then discretized.

For this band gap calculation, the N-Order FDTD algorithm is used [18, 19]. This basis of this algorithm is quite simple: a signal is injected to excite all possible frequencies of the structure. This signal is introduced in accordance to the Maxwell-Gauss law ($div(\vec{E}) = 0$) and given as follows:

$$\vec{E} = \sum_G (\vec{v} \wedge (\vec{k} + \vec{G}) \exp(i(\vec{k} + \vec{G}) \cdot \vec{r}) \quad (9.39)$$

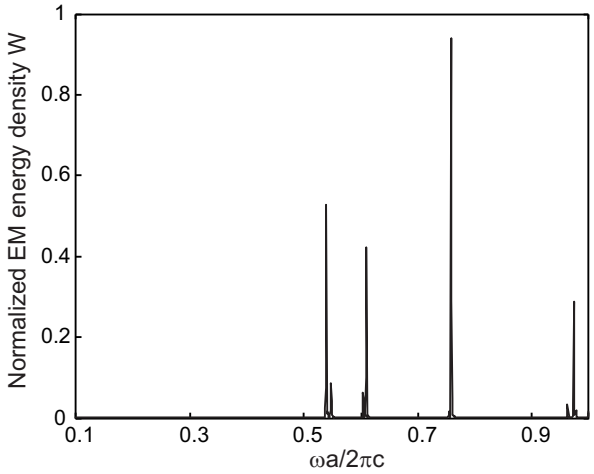


Figure 9.9: Normalized electromagnetic energy density at Γ *point for triangular structure of air holes (of radius* $r = 0.25$a*) into lithium niobate. TM Polarization.*

$\overrightarrow{v}$ is a random vector, $\overrightarrow{k}$ and $\overrightarrow{G}$ denote the wavevector and the reciprocal lattice vector respectively.

After injecting this last initial signal, and for a given $\overrightarrow{k}$, the FDTD simulation is run and electromagnetic energy density time-evolution is calculated as a function of the frequency. This later is calculated through:

$$W = \frac{1}{4}(\varepsilon_0 \varepsilon |E|^2 + \mu |H|^2) \tag{9.40}$$

Only eigenmodes of the structure persist and evanescent ones gradually disappear. After a large number of time iterations (typically 10^5) a permanent regime is then reached and the electromagnetic energy density spectrum exhibits several peaks corresponding to the eigenfrequencies of the studied structure. An example of eigenfrequencies calculation for a triangular structure in the Γ point is shown in figure 9.9. The structure is made of air holes ($n_1 = 1$) into a dielectric medium which is lithium niobate ($LiNbO_3$) with refractive index $n_2 = 2.1421$. The radius of the holes is $r = 0.25$a which corresponds to a filling factor of 0.2267%. The FDTD grid, one PhC period, contains 60×52 spatial grids. To satisfy the stability criterion and avoid numerical dispersion, the time step is taken as $\triangle t = a/(120 \cdot c)$.

To get the complete photonic band structure, it is necessary to scan the k values over all the contour of the irreducible Brillouin zone (ΓXM). Figure 9.10 shows the photonic band diagram calculated for both TE and TM polarizations for a structure parameters similar to those used above in the case of figure 9.9.

We can note the emergence of a photonic bandgap for $\omega a/2\pi c$ between 0.32 and 0.35 in the case of the TE polarization (figure 9.10-a). This band does not exist in the case of the TM polarization (figure 9.10-b) so it is called "partial".

Note here that, for a dispersive material, the calculation of the electromagnetic energy density is no more given by equation 9.40 that is only valid for dielectrics (no dispersion). In the case of metallic dispersive material, the electromagnetic energy density is given by (no magnetic dispersion):

$$W = \frac{1}{4}(\frac{\partial(\omega \varepsilon_0 \varepsilon)}{\partial \omega}|E|^2 + \mu |H|^2) \tag{9.41}$$

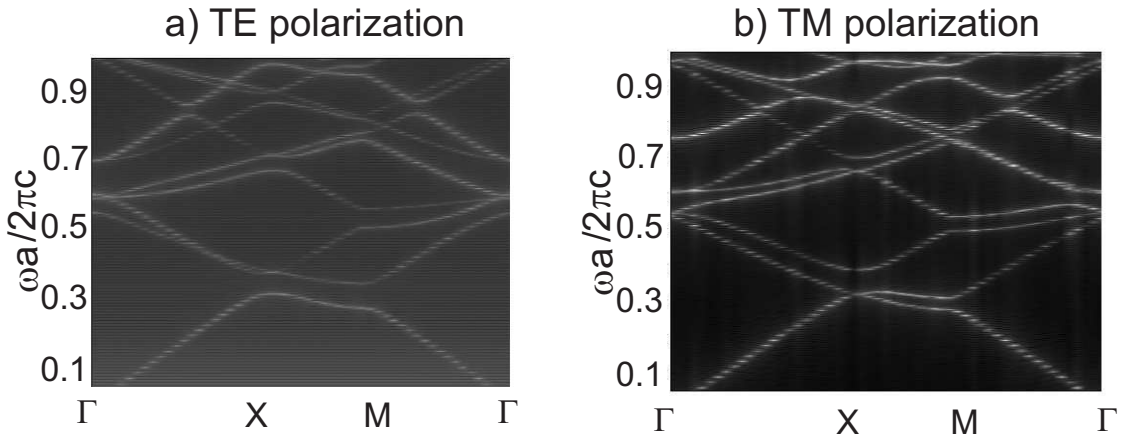


Figure 9.10: Photonic band diagram for triangular structure of air holes (of radius $r = 0.25\,\mathrm{a}$) into lithium niobate.

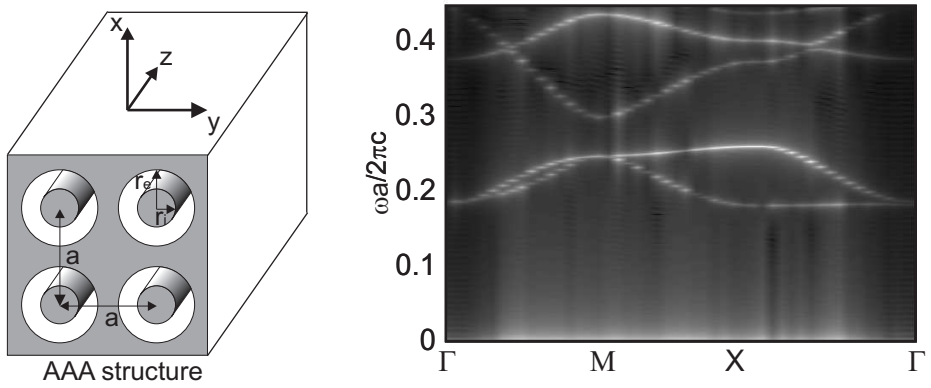


Figure 9.11: In-plane photonic band diagram for annular aperture arrays engraved into silver (TE polarization).

The calculation of the energy density depends then on the dispersion model introduced in the FDTD. Accordingly, an analytic expression of W is obtained through the calculation of the frequency derivative in equation 9.41. Its numerical value is then performed by the determination of the spectral responses of both the two electric and magnetic fields that are determined by the FDTD code.

Another example of band diagram, corresponding to a metallic structure made of annular aperture arrays (AAA) engraved into silver layer and arranged in a square lattice, is shown in figure 9.11. The AAA structure has been proposed by F. Baida and D. Van Labeke [20] for Enhanced Optical Transmission (EOT) applications. It was showed that transmission through AAA sub-wavelength structure could reach 90% in the visible range [21]. This EOT is due to the excitation and the propagation of a guided mode inside each aperture. The main transmission peak corresponds to the excitation of the TE_{11} mode at its cutoff wavelength [19]. This later only depends on the value of the inner and the outer radii. For $r_i = 50\,\mathrm{nm}$ and $r_e = 75\,\mathrm{nm}$ and a lattice constant of $a = 160\,\mathrm{nm}$ one gets the band diagram of figure 9.11.

In case of the figure 9.11, corresponding to the TE polarization, we note the presence of two photonic bandgaps. the first is ranging from zero frequency (infinite wavelength) to the frequency value of $0.1835(c/a)$ ($\lambda = 872\,\mathrm{nm}$). The second gap is in the visible range between 492 nm and 630 nm. Note that these bandgaps are "total" since the corresponding eigen

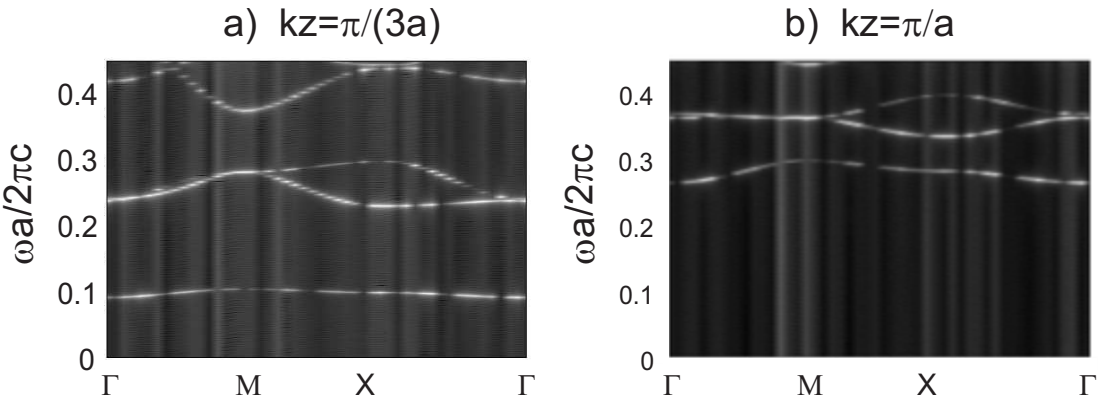


Figure 9.12: Off-plane photonic band diagram for annular aperture arrays made in silver.

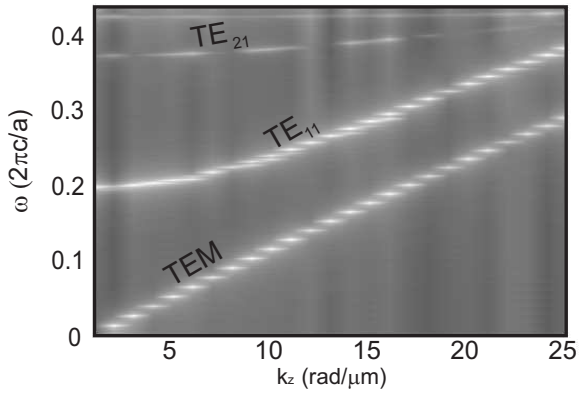


Figure 9.13: Dispersion curves at Γ *point for the coaxial structure made in silver (lattice constante* $a = 160\,\text{nm}$*, inner radius* $r_i = 50\,\text{nm}$ *and outer radius* $r_e = 75\,\text{nm}$*; silver dispersion is modeled by a Drude model).*

frequencies of TM polarization are located above $0.45 \times c/a$.

The figure 9.12 illustrates photonic band diagrams for the same considered AAA structure but in the case of off-plane propagation with two different values of k_z. There is occurrence of an additional photonic band relative to the in-plane case. This is due to the transverse electromagnetic (TEM) mode excited now at a nonzero frequency (far from the cutoff). For $k_z = \pi/(3a)$, the bandgaps are located in the ranges $]1873\,\text{nm}, \infty[$, $]723\,\text{nm}, 1668\,\text{nm}[$ and $]458\,\text{nm}, 575\,\text{nm}[$. These bandgaps become$]653\,\text{nm}, \infty[$, $]512\,\text{nm}, 574\,\text{nm}[$ and $]378\,\text{nm}, 431\,\text{nm}[$ when $k_z = \pi/a$. According to the theory, this band gap shift is due to the fact that the eigenfrequencies of guided modes increase with k_z.

Figure 9.13, showing the dispersion curves (at Γ point) of the guided modes depending on k_z, clearly confirms the TEM nature of the additional mode excited in the off-plan case. This mode band starts from zero frequency, and therefore has no cutoff frequency. An EOT based on the excitation of this peculiar mode can be obtained under two conditions: an oblique incidence with TM polarization [22]. The last section of this chapter is devoted to the study of EOT obtained through the excitation of this peculiar mode.

An example of time evolution of the electromagnetic energy density is given on figure 9.14. The considered structure is an array of coaxial waveguides made in perfectly electric conductor (PEC). All the geometrical parameters are given in the caption in addition to the FDTD simulation ones. One notes that the main peak corresponds to the TE_{21} guided mode

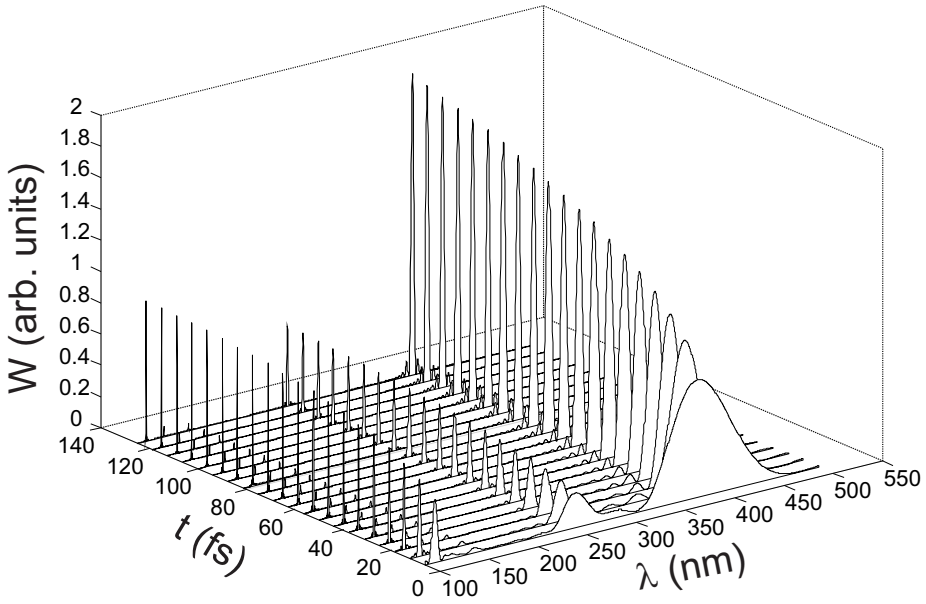


Figure 9.14: Time evolution of the electromagnetic energy density spectrum. The modeled structure is an array of coaxial waveguides made in perfectly electric conductor (PEC) and arranged in square lattice. The inner and outer radii is $r_i = 100nm$ and $r_e = 140nm$ respectively. The period of the grating is $a = 300nm$ but the obtained results are independent on this value because there is non coupling between tow adjacent waveguides. The FDTD simulations are done with a uniform spatial mesh of $\Delta x = \Delta y = \frac{a}{400}$ and the temporal step was fixed to $\Delta t = \frac{\Delta x}{4c}$ where c is the light velocity in vacuum.

that has a cutoff wavelength of $\lambda^c_{TE_{21}} = \frac{\pi(r_i+r_e)}{2}$.

9.3 Scattering calculation for 3D biperiodic nanostructures

In this section, we will focus on the FDTD modeling of dielectric and metallic bi-periodic structures. For normal incidence, the FDTD method, based on the classical Yee's scheme, is a powerful tool that can simply model such periodic structures [24, 25, 26]. In fact, in this simple case, the Floquet-Bloch periodic boundary conditions (PBC) can be easily applied without any change because these conditions are independent of the frequency. However, at oblique incidence, applying PBC implicitly involves a frequency term that must be integrated into the FDTD algorithm that operates in the temporel domain. Thus, in order to adapt FDTD to oblique incidence case, Veysoglu [27] introduced the field transformation method applied to $\overrightarrow{E}$ and $\overrightarrow{H}$ toward new $\overrightarrow{P}$ and $\overrightarrow{Q}$ fields. By the way, the PBC conditions become similar to the ones of normal incidence case nevertheless the immediate consequence of this transformation is the need to modify the Yee's scheme. Several techniques of implementation are then proposed including that of Split-Field Method (SFM) [28].

In the following, we present the reformulation of the FDTD method, based on this SFM technique to adapt it to the case of any incidence. Maxwell's equations are modified and expressed with $\overrightarrow{P}$ and $\overrightarrow{Q}$ variables. They are then discretized using SFM technique. To avoid reflections at the edges of the computational window, the equations in the Berenger's PML medium are also modified and expressed in the new domain within the SFM technique. In addition, the dispersion models mentioned above (Drude, Drude-Lorentz and Drude Critical point models) are also described by modifying and adapting them to the SFM technique.

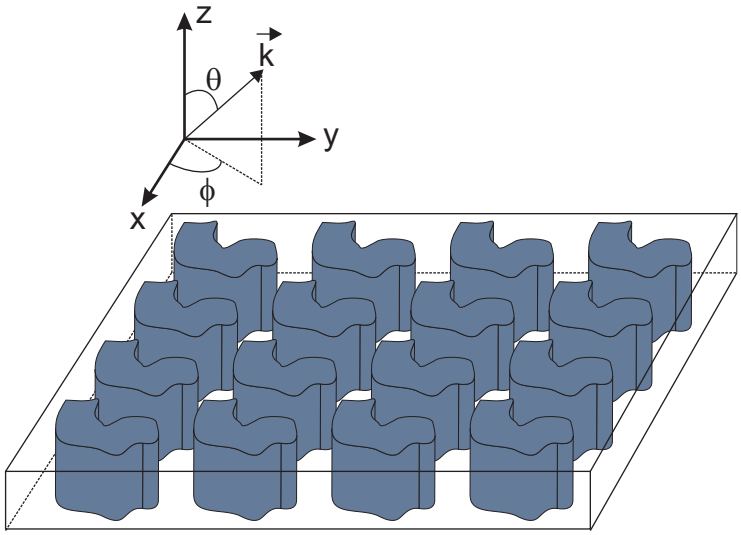


Figure 9.15: Sketch of the biperiodic structure illuminated by plane wave propagating along the $\overrightarrow{k}$ vector defined by its Euler angles θ and ϕ.

9.3.1 Position of the problem: New $\overrightarrow{P} - \overrightarrow{Q}$ variables

Let us consider a bi-periodic structure, finished along the third direction and illuminated by a plane wave propagating at oblique incidence (see figure 9.15).

According to the notations of figure 9.15, the electric and magnetic fields of the incident plane wave can be written as:

$$\overrightarrow{E}_i = \overrightarrow{E}_{0i} e^{i\left[k_x \cdot x + k_y \cdot y + k_z \cdot z + \omega \cdot t\right]} \tag{9.42.a}$$

$$\overrightarrow{H}_i = \overrightarrow{H}_{0i} e^{i\left[k_x \cdot x + k_y \cdot y + k_z \cdot z + \omega \cdot t\right]} \tag{9.42.b}$$

where:

$$k_x = \frac{\omega}{v} \sin\theta \cos\varphi \tag{9.43}$$

$$k_y = \frac{\omega}{v} \sin\theta \sin\varphi \tag{9.44}$$

$$k_z = \frac{\omega}{v} \cos\theta \tag{9.45}$$

For the periodic object, a single pattern (one period) is then considered for the FDTD calculation (see figure 9.6). The periodic conditions are then written so that the fields on one side of the calculation window are expressed versus the fields on the opposite side through the Floquet-Bloch conditions. For x (lattice constant a) and y (lattice constant b) periodic structures, these conditions are expressed as follows:

$$\overrightarrow{E}(x, y, z, t) = \overrightarrow{E}(x+a, y, z, t) \cdot e^{-ik_x \cdot a} \tag{9.46.a}$$

$$\overrightarrow{E}(x, y, z, t) = \overrightarrow{E}(x, y+b, z, t) \cdot e^{-ik_y \cdot b} \tag{9.46.b}$$

$$\overrightarrow{H}(x+a, y, z, t) = \overrightarrow{H}(x, y, z, t) \cdot e^{ik_x \cdot a} \tag{9.46.c}$$

$$\overrightarrow{H}(x, y+b, z, t) = \overrightarrow{H}(x, y, z, t) \cdot e^{ik_y \cdot b} \tag{9.46.d}$$

As the FDTD method operates in the temporal domain and k_x and k_y components explicitly depend of ω, the direct application of these periodic conditions is prohibited. Consequently, a change of variables is performed so that $\overrightarrow{E}$ and $\overrightarrow{H}$ components are replaced by two new components $\overrightarrow{P}$ and $\overrightarrow{Q}$ respectively in order to eliminate the k_x and k_y dependence in the PBC. These new fields are defined as follows:

$$\overrightarrow{P} = \overrightarrow{E} \cdot e^{-ik_x x} \cdot e^{-ik_y y} \quad (9.47.a)$$
$$\overrightarrow{Q} = \overrightarrow{H} \cdot e^{-ik_x x} \cdot e^{-ik_y y} \quad (9.47.b)$$

Therefore, the new periodic conditions can be applied similarly to the case of normal incidence through the relations:

$$\overrightarrow{P}(x, y, z, t) = \overrightarrow{P}(x+a, y, z, t) \quad (9.48.a)$$
$$\overrightarrow{Q}(x+a, y, z, t) = \overrightarrow{Q}(x, y, z, t) \quad (9.48.b)$$
$$\overrightarrow{P}(x, y, z, t) = \overrightarrow{P}(x, y+b, z, t) \quad (9.48.c)$$
$$\overrightarrow{Q}(x, y+b, z, t) = \overrightarrow{Q}(x, y, z, t) \quad (9.48.d)$$

Replacing $\overrightarrow{E}$ and $\overrightarrow{H}$ by their expressions in terms of $\overrightarrow{P}$ and $\overrightarrow{Q}$ through equations 9.47 in Maxwell's equations system 9.5 leads to:

$$\frac{\partial Q_x}{\partial t} = \frac{1}{\mu_0}\left[\frac{\partial P_y}{\partial z} - \frac{\partial P_z}{\partial y} - ik_y P_z\right] \quad (9.49.a)$$
$$\frac{\partial Q_y}{\partial t} = \frac{1}{\mu_0}\left[\frac{\partial P_z}{\partial x} + ik_x P_z - \frac{\partial P_x}{\partial z}\right] \quad (9.49.b)$$
$$\frac{\partial Q_z}{\partial t} = \frac{1}{\mu_0}\left[\frac{\partial P_x}{\partial y} + ik_y P_x - \frac{\partial P_y}{\partial x} - ik_x P_y\right] \quad (9.49.c)$$
$$\frac{\partial P_x}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_z}{\partial y} + ik_y Q_z - \frac{\partial Q_y}{\partial z}\right] \quad (9.49.d)$$
$$\frac{\partial P_y}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_x}{\partial z} - \frac{\partial Q_z}{\partial x} - ik_x Q_z\right] \quad (9.49.e)$$
$$\frac{\partial P_z}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_y}{\partial x} + ik_x Q_y - \frac{\partial Q_x}{\partial y} - ik_y Q_x\right] \quad (9.49.f)$$

We can notice that for a wave propagating at normal incidence, the system (9.49) above is equivalent to the conventional Maxwell' equations expressed in $\overrightarrow{E} - \overrightarrow{H}$. Nonetheless, in the oblique case, additional terms appear in the second right members of equations (9.49) and they explicitly depend on k_x and k_y i.e. on the frequency ω. Even if these terms are equivalent to time derivatives, the direct implementation of the FDTD in this case is impossible. Many implementation techniques have been proposed [29, 30, 31, 28, 32, 33] to overcome this problem. One of them is the Split Field Method [32, 28] which will be described below.

9.3.2 Split Field Method

SFM technique is based on the split of $\overrightarrow{P}$ and $\overrightarrow{Q}$ field components. To illustrate the method, let us take for example the split of the Q_x component occurring in equation (9.49.a). By reducing the frequency additional term on the left hand, this equation can be written as:

$$\frac{\partial Q_x}{\partial t} + i\omega \frac{k_y}{\mu\omega} P_z = \frac{1}{\mu}\left[\frac{\partial P_y}{\partial z} - \frac{\partial P_z}{\partial y}\right] \tag{9.50a}$$

According to (9.42.a) and (9.47.a), equation (9.50a) becomes:

$$\frac{\partial}{\partial t}\left[Q_x + \frac{k_y}{\mu\omega} P_z\right] = \frac{1}{\mu}\left[\frac{\partial P_y}{\partial z} - \frac{\partial P_z}{\partial y}\right] \tag{9.51a}$$

This leads to a new component $Q_{xa} = Q_x + \frac{k_y}{\mu\omega} P_z$ which satisfies Maxwell's equation as for normal incidence. Similarly, the split of all the others components in the $\overrightarrow{P} - \overrightarrow{Q}$ domain gives:

$$Q_{xa} = Q_x + \frac{k_y}{\mu\omega} P_z \tag{9.52.a}$$

$$Q_{ya} = Q_y - \frac{k_x}{\mu\omega} P_z \tag{9.52.b}$$

$$Q_{za} = Q_z - \frac{k_y}{\mu\omega} P_x + \frac{k_x}{\mu\omega} P_y \tag{9.52.c}$$

$$P_{xa} = P_x - \frac{k_y}{\varepsilon\omega} Q_z \tag{9.52.d}$$

$$P_{ya} = P_y + \frac{k_x}{\varepsilon\omega} Q_z \tag{9.52.e}$$

$$P_{za} = P_z - \frac{k_x}{\varepsilon\omega} Q_y + \frac{k_y}{\varepsilon\omega} Q_x \tag{9.52.f}$$

The six components thereby obtained satisfy the following equations that can be discretized according to the classical Yee's scheme:

$$\frac{\partial Q_{xa}}{\partial t} = \frac{1}{\mu}\left[\frac{\partial P_y}{\partial z} - \frac{\partial P_z}{\partial y}\right] \tag{9.53.a}$$

$$\frac{\partial Q_{ya}}{\partial t} = \frac{1}{\mu}\left[\frac{\partial P_z}{\partial x} - \frac{\partial P_x}{\partial z}\right] \tag{9.53.b}$$

$$\frac{\partial Q_{za}}{\partial t} = \frac{1}{\mu}\left[\frac{\partial P_x}{\partial y} - \frac{\partial P_y}{\partial x}\right] \tag{9.53.c}$$

$$\frac{\partial P_{xa}}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_z}{\partial y} - \frac{\partial Q_y}{\partial z}\right] \tag{9.53.d}$$

$$\frac{\partial P_{ya}}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_x}{\partial z} - \frac{\partial Q_z}{\partial x}\right] \tag{9.53.e}$$

$$\frac{\partial P_{za}}{\partial t} = \frac{1}{\varepsilon}\left[\frac{\partial Q_y}{\partial x} - \frac{\partial Q_x}{\partial y}\right] \tag{9.53.f}$$

Once the updated components of $\overrightarrow{P}_a$ and $\overrightarrow{Q}_a$ completed, the second step of the algorithm is to calculate $\overrightarrow{P}$ and $\overrightarrow{Q}$ components through the system of equations (9.52) that gives after simple algebra the system below:

$$Q_z = \frac{1}{1-\frac{k_x^2+k_y^2}{\varepsilon\mu\omega^2}}\left[Q_{za} + \frac{k_y}{\mu\omega}P_{xa} - \frac{k_x}{\mu\omega}P_{ya}\right] \tag{9.54.a}$$

$$P_z = \frac{1}{1-\frac{k_x^2+k_y^2}{\varepsilon\mu\omega^2}}\left[P_{za} + \frac{k_x}{\varepsilon\omega}Q_{ya} - \frac{k_y}{\varepsilon\omega}Q_{xa}\right] \tag{9.54.b}$$

$$Q_x = Q_{xa} - \frac{k_y}{\mu\omega}P_z \tag{9.54.c}$$

$$Q_y = Q_{ya} + \frac{k_x}{\mu\omega}P_z \tag{9.54.d}$$

$$P_x = P_{xa} + \frac{k_y}{\varepsilon\omega}Q_z \tag{9.54.e}$$

$$P_y = P_{ya} - \frac{k_x}{\varepsilon\omega}Q_z \tag{9.54.f}$$

This system (9.54) needs to calculate $\overrightarrow{P}$ and $\overrightarrow{Q}$ components at the same time iteration as $\overrightarrow{P}_a$ and $\overrightarrow{Q}_a$ components. This is in contradiction with the traditional Yee's scheme. Consequently, the new ($\overrightarrow{P}$, $\overrightarrow{Q}$) and ($\overrightarrow{P}_a$, $\overrightarrow{Q}_a$) fields will be calculated at time $n\Delta t$ and time $(n+\frac{1}{2})\Delta t$ in order to reach a stable numerical schema. To this end, each component is calculated twice in one time iteration by introducing other intermediate components in the calculation program (see reference [34]).

Stability criterion

As the transition to the new $\overrightarrow{P} - \overrightarrow{Q}$ domain, the stability criterion is also modified. Based on the calculation of Kao [29, 30] and in the case of 3D uniform meshing, this later is expressed as:

$$\frac{\Delta}{\Delta t} \geq \frac{v_i}{v_i^2 \mu\varepsilon - \sin^2(\theta)} \left\{ |\sin(\theta)\cdot\cos(\varphi)| + |\sin(\theta)\cdot\sin(\varphi)| + \sqrt{3v_i^2\mu\varepsilon - 2\cdot\sin^2(\theta)\left(1 - |\sin(\varphi)\cdot\cos(\varphi)|\right)} \right\} \quad (9.55)$$

where v_i is the phase velocity of the incident wave and ε and μ are chosen to be the characteristics of the less dense medium in the computational domain.

Let us note here that the time step decreases with the incidence angle θ and hence the computational time becomes very long for large incidence angles. Nonetheless, the computational time is relatively acceptable up to an incidence angle of 80^o.

9.3.3 Absorbing boundary conditions : PML

The implementation of absorbing boundary conditions in the oblique case requires to make a change of variables on the fields components in the PML medium similarly to the changes made in the main computational grid [34]. For x and y periodic structure, only PML is needed in the third direction (Oz). In this case, the new fields components are defined as follows:

$$P_{\nu\mu} = E_{\nu\mu}\cdot e^{-ik_x x}\cdot e^{-ik_y y} \quad (9.56.a)$$

$$Q_{\nu\mu} = H_{\nu\mu}\cdot e^{-ik_x x}\cdot e^{-ik_y y} \quad (9.56.b)$$

$$P_z = E_z\cdot e^{-ik_x x}\cdot e^{-ik_y y} \quad (9.56.c)$$

where ν represents x or y and μ denotes x, y or z. $E_{\nu\mu}$ and $Q_{\nu\mu}$ are the field components in the classical PML shell corresponding to the components of the two fictitious waves resulting from the split of the plane wave inside the PML (see section 1 of this chapter). For details of implementing these PML at oblique incidence, the reader can refer to [34].

9.3.4 SFM-FDTD in dispersive media

For oblique incidence, and according to the two systems of equations (9.53) and (9.54), the components that require particular treatment in the dispersive medium are: P_{xa}, P_{ya}, P_{za}, Q_z, P_z, P_x and P_y. Direct calculation of these components by equations (9.53) and (9.54) involves the permittivity term which is frequency-dependent. In this section, we only show how to take into account the media dispersion in FDTD oblique incidence in the case of the Drude critical points model [35]. The implementation details of the other of dispersion models by SFM-FDTD are given in [36] for Debye model, and in [37] for both Drude and Drude-Lorentz models.

Let us quote that equations (9.53) for the calculation of P_{xa}, P_{ya} and P_{za} are similar to traditional Maxwell's equations. Accordingly, the calculation of these components in the dispersive medium will not require any further treatment compared to the normal incidence case. Contrarily, equations (9.54) for the Q_z, P_z, P_x and P_y need a different way to be processed.

P_{xa}, P_{ya} and P_{za} implementation: These three components are calculated in a similar way. So let us take as an example only the P_{xa} calculation. By analogy with the normal incidence case (equation 9.25), we introduced a new component L_{xa} (equivalent to the D_x component in the classical case) defined as:

$$L_{xa} = \varepsilon_0 \cdot \varepsilon_{DCP} \cdot P_{xa} \tag{9.57}$$

Equation (9.53.d) is therefore wrote as:

$$\frac{\partial L_{xa}}{\partial t} = \left[\frac{\partial Q_z}{\partial y} - \frac{\partial Q_y}{\partial z}\right] \tag{9.58}$$

The discretization of this last equation allows us to calculate the L_{xa} variable as follows:

$$L_{xa}^{n+1}{}_{\left(i+\frac{1}{2},j,k\right)} = L_{xa}^{n}{}_{\left(i+\frac{1}{2},j,k\right)} + \frac{\Delta t}{\Delta y}\left[Q_z^n{}_{\left(i+\frac{1}{2},j+\frac{1}{2},k\right)} - Q_z^n{}_{\left(i+\frac{1}{2},j-\frac{1}{2},k\right)}\right] + \frac{\Delta t}{\Delta z}\left[Q_y^n{}_{\left(i+\frac{1}{2},j,k-\frac{1}{2}\right)} - Q_y^n{}_{\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}\right] \tag{9.59}$$

Analogically to equations (9.26), (9.27.a) and (9.27.b), L_{xa} can be expressed as follows:

$$L_{xa} = L_{xa_D} + \sum_{p=1}^{p=2} L_{xa_{C_p}} \tag{9.60}$$

with:

$$L_{xa_D} = \varepsilon_0[\varepsilon_\infty - \frac{\omega_P^2}{\omega^2 + \mathrm{i}\gamma\omega}]P_{xa} \tag{9.61.a}$$

$$L_{xa_{C_p}} = \varepsilon_0[A_p\Omega_p(\frac{\mathrm{e}^{\mathrm{i}\phi_p}}{\Omega_p - \omega - \mathrm{i}\Gamma_p} + \frac{\mathrm{e}^{-\mathrm{i}\phi_p}}{\Omega_p + \omega + \mathrm{i}\Gamma_p})]P_{xa} \tag{9.61.b}$$

As before, after the inverse Fourier transforms and finite centred differences discretization of different partial derivatives, we reach the updated equations for the component P_{xa} :

$$P_{xa}^{n+1} = \frac{1}{\frac{\chi_D}{\alpha_D} + \sum_{p=1}^{p=2}(\frac{\chi_p}{\alpha_p})}\left[L_{xa}^{n+1} + \frac{\beta_D}{\alpha_D}L_{xa_D}^{n-1} + \frac{4}{\alpha_D}L_{xa_D}^{n} - \frac{\delta_D}{\alpha_D}P_{xa}^{n-1} - \frac{4\varepsilon_0\varepsilon_\infty}{\alpha_D}P_{xa}^{n} + \sum_{p=1}^{p=2}(\frac{\beta_p}{\alpha_p}L_{xa_{C_p}}^{n-1} - \frac{4}{\alpha_p}L_{xa_{C_p}}^{n}) + \sum_{p=1}^{p=2}(\frac{\delta_p}{\alpha_p})P_{xa}^{n-1}\right] \tag{9.62.a}$$

$$L_{xa_D}^{n+1} = \frac{1}{\alpha_D}\left[-\beta_D L_{xa_D}^{n-1} - 4L_{xa_D}^{n} + \chi_D P_{xa}^{n+1} + \delta_D P_{xa}^{n-1} + 4\varepsilon_0\varepsilon_\infty P_{xa}^{n}\right] \tag{9.62.b}$$

$$L_{xa_{C_p}}^{n+1} = \frac{1}{\alpha_p}[-\beta_p L_{xa_{C_p}}^{n-1} + 4L_{xa_{C_p}}^{n} + \chi_p P_{xa}^{n+1} + \delta_p P_{xa}^{n-1}] \tag{9.62.c}$$

Q_z, P_z, P_x and P_y implementation: The calculation of the remaining components Q_z, P_z, P_x and P_y needs the introduction of other variables involving other equations. We consider as an

example the P_z component for which implementation equations are detailed. Equation (9.54.b) involves the following one:

$$\varepsilon \cdot M_z = \frac{k_x}{\omega} Q_{ya} - \frac{k_y}{\omega} Q_{xa} + \frac{k_x^2 + k_y^2}{\mu \omega^2} P_z \tag{9.63}$$

with:

$$M_z = P_z - P_{za} \tag{9.64}$$

By setting:

$$T_z = \frac{k_x}{\omega} Q_{ya} - \frac{k_y}{\omega} Q_{xa} + \frac{k_x^2 + k_y^2}{\mu \omega^2} P_z \tag{9.65}$$

equation (9.63) becomes:

$$T_z = \varepsilon \cdot M_z = \varepsilon_0 \varepsilon_{DCP} M_z \tag{9.66}$$

As considered above, the T_z component can be expressed as:

$$T_z = T_{z_D} + \sum_{p=1}^{p=2} T_{z_{Cp}} \tag{9.67}$$

with:

$$T_{z_D} = \varepsilon_0 [\varepsilon_\infty - \frac{\omega_P^2}{\omega^2 + \mathrm{i}\gamma\omega}] M_z \tag{9.68.a}$$

$$T_{z_{C_p}} = \varepsilon_0 [A_p \Omega_p (\frac{\mathrm{e}^{\mathrm{i}\phi_p}}{\Omega_p - \omega - \mathrm{i}\Gamma_p} + \frac{\mathrm{e}^{-\mathrm{i}\phi_p}}{\Omega_p + \omega + \mathrm{i}\Gamma_p})] M_z \tag{9.68.b}$$

Based on the inverse Fourier transforms of the equations (9.68) above, centered difference approximations for the derivatives and taking into account the equations (9.65), (9.67) and (9.66), we get:

$$\begin{aligned} M_z^{n+1} = \frac{1}{\frac{\chi_D}{\alpha_D} + \sum_{p=1}^{p=2} (\frac{\chi_p}{\alpha_p}) - \frac{k_x^2 + k_y^2}{\mu \omega^2}} \Bigg[& \frac{k_x^2 + k_y^2}{\mu \omega^2} P_z^{n+1} + \frac{k_x}{\omega} Q_{ya}^{n+1} - \frac{k_y}{\omega} Q_{xa}^{n+1} + \frac{\beta_D}{\alpha_D} T_{z_D}^{n-1} + \frac{4}{\alpha_D} T_{z_D}^n \\ & - \frac{\delta_D}{\alpha_D} M_z^{n-1} - \frac{4\varepsilon_0 \varepsilon_\infty}{\alpha_D} M_z^n + \sum_{p=1}^{p=2} (\frac{\beta_p}{\alpha_p} T_{z_{Cp}}^{n-1} - \frac{4}{\alpha_p} T_{z_{Cp}}^n) + \sum_{p=1}^{p=2} (\frac{\delta_p}{\alpha_p}) M_z^{n-1} \Bigg] \end{aligned} \tag{9.69.a}$$

$$T_{z_D}^{n+1} = \frac{1}{\alpha_D} \left[-\beta_D T_{z_D}^{n-1} - 4 T_{z_D}^n + \chi_D M_z^{n+1} + \delta_D M_z^{n-1} + 4\varepsilon_0 \varepsilon_\infty M_z^n \right] \tag{9.69.b}$$

$$T_{z_{Cp}}^{n+1} = \frac{1}{\alpha_p} [-\beta_p T_{z_{Cp}}^{n-1} + 4 T_{z_{Cp}}^n + \chi_p M_z^{n+1} + \delta_p M_z^{n-1}] \tag{9.69.c}$$

$$P_z^{n+1} = M_z^{n+1} + P_{za}^{n+1} \tag{9.69.d}$$

The equations to update the Q_z, P_x and P_y components are obtained by the same process.

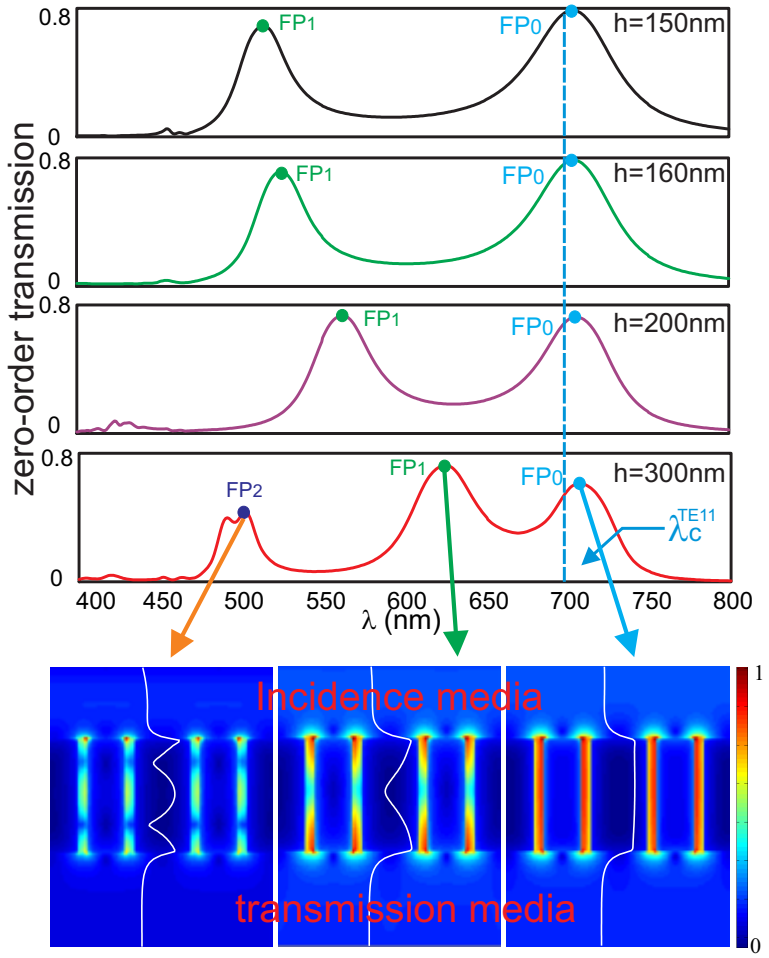


Figure 9.16: Up: Transmission spectra at normal incidence of an AAA structure made in silver film with different thicknesses values (H). The geometrical parameters of the annular apertures are $r_e = 75$ nm,$r_i = 50$ nm and the period is fixed to $a = 300$ nm. Down: Electric field intensity distributions around the apertures showing the interference patterns that take place inside them along the metal thickness direction. For FP_0 peak, the TE_{11} guided mode is excited at its cutoff wavelength so that the phase velocity tends to infinity and the effective index falls to zero. In this case, EOT occurs whatever is the value of the thickness because the phase matching condition is automatically fulfilled.

9.3.5 3D-SFM-FDTD application: EOT at oblique incidence through AAA structures

Let us recall the origin of the EOT through the AAA structure: as mentioned before, at normal incidence it is only due to the excitation of the TE_{11} guided mode inside each annular aperture. In this case, the obtained EOT is angle and polarization-independent and its spectral position corresponds to the cutoff wavelength of this guided mode. Consequently, it does not depends either on the metal thickness even if some additional peaks appear in the transmission spectrum when the thickness increases (see figure 9.16).

These peaks (named FP_m, $m \in \Re$ on figure 9.16) are Fabry-Perot harmonics of the TE_{11} mode that occur at fixed values of the wavelength fulfilling a phase matching condition:

$$\lambda_{TE_{11}}(m\pi - \phi_r) = 2\pi n_{eff}^{TE_{11}} H \tag{9.70}$$

where $n_{eff}^{TE_{11}}$ is the real part of the effective index of the guided mode, ϕ_r is the phase

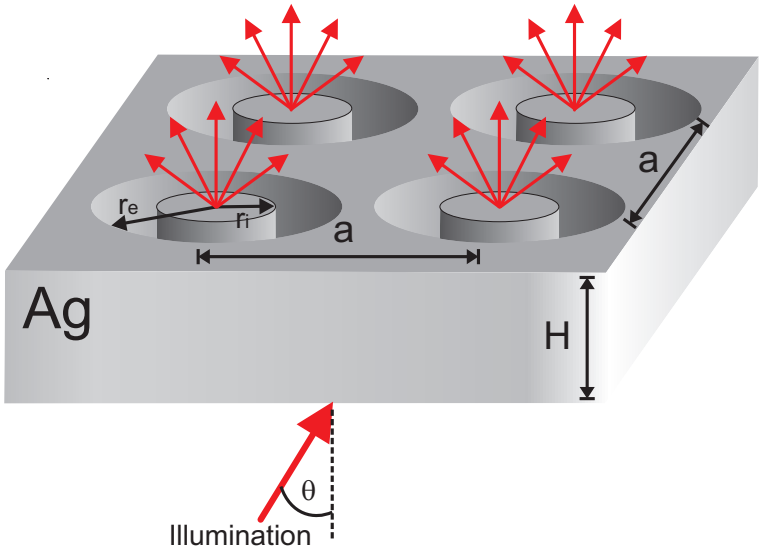


Figure 9.17: Schematic of a classical annular aperture array (AAA). r_e is the outer radius, r_i is the inner one, a is the period and θ is the angle of incidence.

change induced by the reflection on the two ends of the annular aperture and H is the metallic film thickness. At the cutoff, the effective index of the guided mode becomes very small leading to a phase matching that does not depend on the metal thickness. Nevertheless, a small spectral shift can appear between the cutoff value and the position of the transmission peak due to $\phi_r \neq 0$. This shift is clearly shown on all the spectra of figure 9.16 but it seems to be more important in the case of thicker plates (here $H = 300\ nm$). In fact, the phase ϕ_r can be seen as the result of the conversion between the incident plane wave and the guide mode through diffraction phenomenon that must depends on the metal thickness.

Let us now consider the case of oblique incidence (see figure 9.17): as mentioned before, EOT can appear through the excitation of both the TE_{11} and the TEM modes. In fact only few papers have discussed on this mode [38, 39] while its excitation conditions were recently analytically derived reference [22].

Indeed, this later is only excited with the TM polarization component of the incident beam. FDTD simulations in the case of both PEC (see figure 9.18) and real dispersive metal (figure 9 of reference [37, 40]) are done and demonstrate the occurrence of additional transmission peaks due to the excitation of the TEM guided mode. Nevertheless, others configurations such as the Slanted AAA (SAAA), that was proposed first by S. Nosal and J.J. Greffet [41], also demonstrate a possible excitation of the TEM mode for any incidence angle including normal incidence.

Moreover, as for the TE_{11} mode, the spectral position of the TEM-transmission peaks is driven by a similar phase matching condition given by equation 9.70. Nonetheless, the zero harmonic (FP_0 for $m = 0$) is now expelled to infinity and only higher orders correspond to a finite value of the wavelength. In this case, the metal thickness becomes a very important parameter that permits to adapt the transmission peak at a desired value of wavelength. Unluckily, only relatively thick metal plates allow the excitation and the propagation of the TEM mode.

Nevertheless, even if the TEM mode is excited in oblique incidence with conventional AAA (see figure 9.20a) or at normal incidence through SAAA (figure 9.20b), the transmission

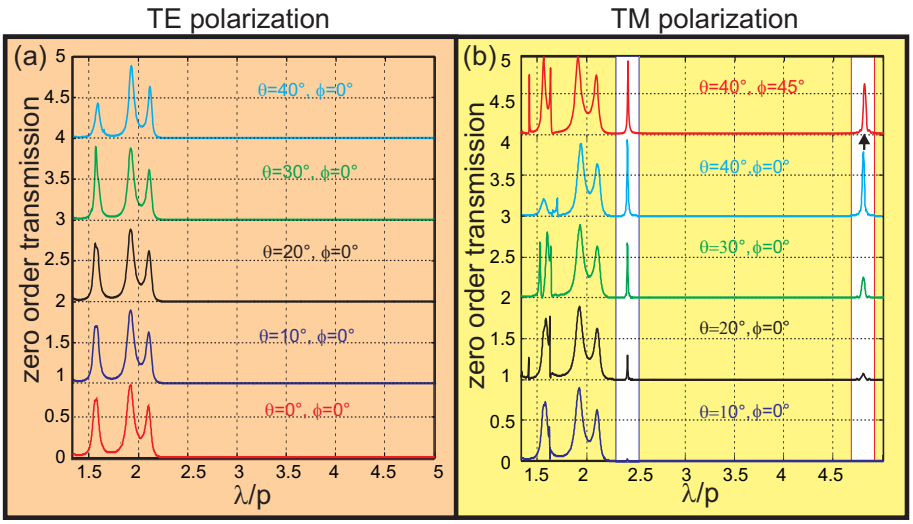


Figure 9.18: Transmission spectra through AAA structure made in perfectly electric conductor and illuminated by a TE (left) and TM (right) linearly polarized plane wave. As depicted on figure 9.15, θ and ϕ denote the incidence and azimuthal angle respectively. The geometrical parameter of the AAA structure are: $r_e = a/3, r_i = a/4$ and $H = 2a$ (please see figure 9.17 for notations). Two families of TEM peak are pointed out using two white vertical rectangles. The right one corresponds here to the first Fabry-Perot harmonic and the left one frames the second harmonic. Note that other higher harmonics also occur at smaller wavelength values.

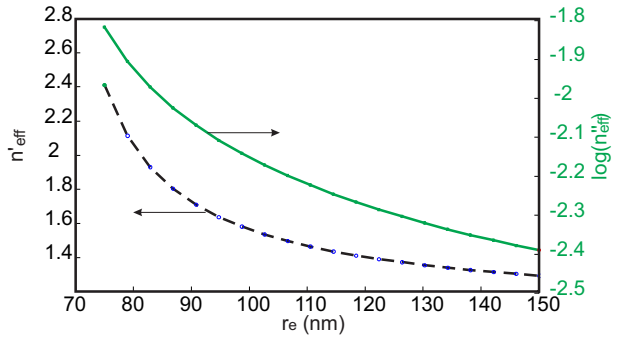


Figure 9.19: Real part n'_{eff} and log_{10} of the imaginary part $n"_{eff}$ of the effective index associated with the TEM-like mode of an infinite coaxial waveguide as a function of the outer radius r_e. The inner radius is set to $r_i = 65$ nm and the working wavelength is $\lambda = 1550$ nm.

efficiency remains very weak with regard to the TE_{11} mode. This is essentially due to metal losses. In fact, and as it can be shown in figure 9.19, the imaginary part of the effective index of the TEM-like guided mode is fairly consistent and can not be negligible.

Fortunately, another solution that is currently used in the radio-frequency domain to increase the impedance adaptation between a coaxial antenna and the vacuum can be envisaged to enhance the transmission coefficient: it consists in stretching out the central metallic part of the coaxial waveguide with respect to the outside electrode. This configuration was implicitly proposed in reference [42] to achieve 90% light transmission thanks to the excitation of the TEM mode. This kind of structure design and fabrication is readily achievable at radio frequencies. Unfortunately, this becomes more difficult in the visible range but remains possible through manufacturing process having nanometric resolution such as new generation of Focused Ion Beam.

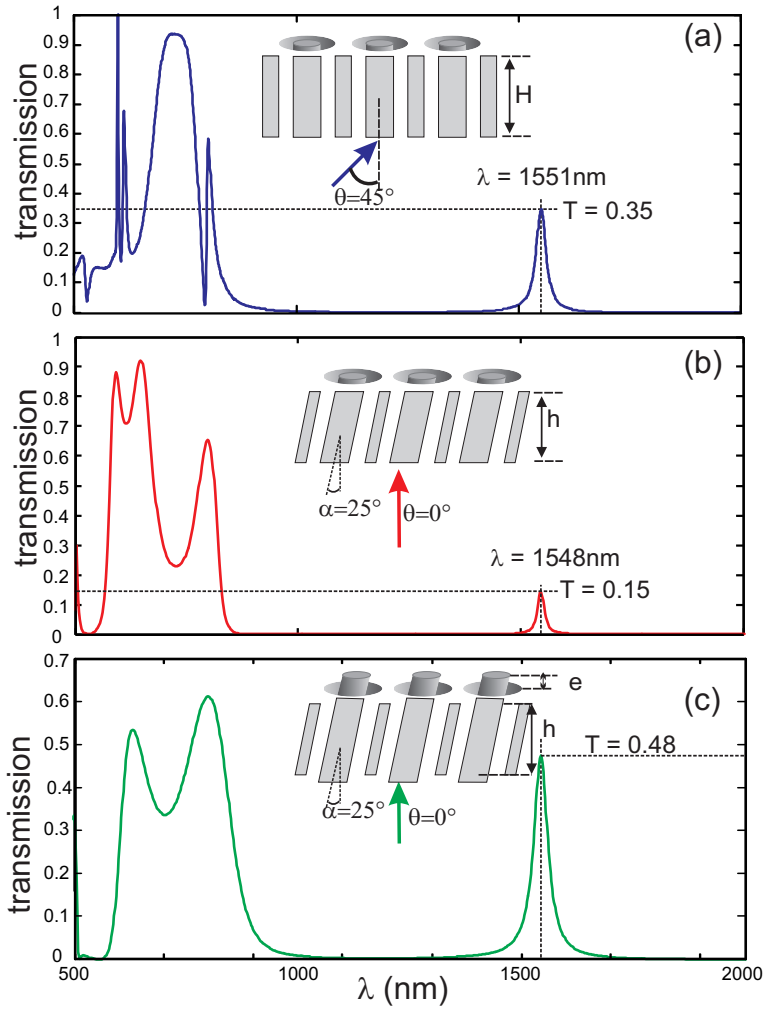


Figure 9.20: Zero-order transmission spectra for three different AAA configurations where outer and inner radii are fixed to $r_e = 130$ nm and $r_i = 65$ nm respectively. (a) Conventional structure illuminated at 45^o (metal thickness of $h = 495$ nm). (b) SAAA with tilt angle of 25^o with respect to the vertical direction. The thickness of the metallic film ($h = 430$ nm) is chosen in order to get a TEM peak transmission at $\lambda = 1550$ nm. (c) SAAA structure with inner metallic parts that stretch out from the metallic film over a distance $e = 80$ nm. This allows increasing of the impedance matching between the in- and out-coming plane waves with the TEM guided mode inside the apertures. The metal thickness is also adjusted to in order to get a TEM peak at $\lambda = 1550$ nm with a transmission efficiency of 48%.

9.4 Conclusion

The FDTD is a powerful tool to model periodic and aperiodic structures. The time evolution of the electromagnetic field is directly evaluated and allows to follow the light propagation inside and around the studied structure. The SFM technique extends the FDTD capabilities to treat the diffraction problem for any incidence angle or any polarization. The integration of dispersion models such as Drude critical point allows accurate simulations that take into account the effective dispersion of noble metals in the considered spectral range especially in the visible domain. Nevertheless, the number of electromagnetic field components grows rapidly and can be larger than 100 in some particular cases (in the PML region with Drude-Lorentz dispersion model for instance). In spite of all these criticisms, the FDTD is actually one of the most used method to model experiments in Nano-Optics as attested by the number of publications in this area.

References:

[1] K. S. Yee. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, 14:302–307, 1966.

[2] A. Taflove and M. E. Brodwin. Computation of the electromagnetic fields and induced temperatures within a model of the microwave-irradiated human eye. *IEEE Transactions on Microwave Theory and Techniques*, 23:888–896, 1975.

[3] A. Taflove and M. E. Brodwin. Numerical solution of steady-state electromagnetic scattering problems using the time-dependent maxwell's equations. *IEEE Transactions on Microwave Theory and Techniques*, 23:623–630, 1975.

[4] K. S. Kunz and K.-M. Lee. A three-dimensional finite-difference solution of the external response of an aircraft to a complex transient em environment: Part i-the method and its implementation. *IEEE Transactions on Electromagnetic Compatibility*, 20:328–333, 1978.

[5] R. Courant, K. O. Friedrich, and H. Lewy. On the partial difference equations of mathematical physics. *IBM Journal of Research and Development*, 11:215–234, 1967.

[6] A. Taflove and S. C. Hagness. *Computational Electrodynamics. The Finite-Difference Time-Domain Method, 2nd ed.* Artech House, Norwood, MA, 2005.

[7] J.-P. Berenger. A PERFECTLY MATCHED LAYER FOR THE ABSORPTION OF ELECTROMAGNETIC-WAVES. *JOURNAL OF COMPUTATIONAL PHYSICS*, 114(2):185–200, 1994.

[8] C. F. Bohren and D. R. Huffman. *Absorption and Scattering of Light by Small Particles*. Wiley-Interscience, New York, 1983.

[9] N. Ashcroft and N. D. Mermin. *Physique des Solides*. EDP Sciences, Les Ulis, 2002.

[10] A. Vial, A. S. Grimault, D. Macias, D. Barchiesi, and M. Lamy de la Chapelle. Improved analytical fit of gold dispersion: application to the modeling of extinction spectra with a finite-difference time-domain method. *Phys. Rev. B*, 71:085416, 2005.

[11] L. Novotny and B. Hecht. *Principles of Nano-optic*. Cambridge University Press, 2006.

[12] P. G. Etchegoin, E. C. Le Ru, and M. Meyer. An analytic model for the optical properties of gold. *The Journal of Chemical Physics*, 125(16):164705, 2006.

[13] P. Etchegoin, J. Kircher, and M. Cardona. Elasto-optical constants of si. *Phys. Rev. B*, 47:10292–10303, 1993.

[14] F. Hao and P. Nordlander. Efficient dielectric function for fdtd simulation of the optical properties of silver and gold nanoparticles. *Chemical Physics Letters*, 446(1-3):115 – 118, 2007.

[15] A. Vial and T. Laroche. Comparison of gold and silver dispersion laws suitable for fdtd simulations. *Applied Physics B: Lasers and Optics*, 93:139–143, 2008.

[16] M. Qiu. *Computational methods for the analysis and design of photonic bandgap structures*. Ph. d., Royal Institute of Technology, Stockholm, 2000.

[17] W. Kuang, W. J. Kim, and J. D. O'Brien. Finite-difference time domain method for nonorthogonal unit-cell two-dimensional photonic crystals. *J. Lightwave Technol.*, 25(9):2612–2617, 2007.

[18] C. T. Chan, Q. L. Yu, and K. M. Ho. Order-n spectral method for electromagnetic waves. *Phys. Rev. B*, 51(23):16635–16642, 1995.

[19] F. I. Baida, D. Van Labeke, G. Granet, A. Moreau, and A. Belkhir. Origin of the super-enhanced light transmission through a 2-d metallic annular aperture array: a study of photonic bands. *Applied Phys. B*, 79(1):1–8, 2004.

[20] F. I. Baida and D. Van Labeke. Light transmission by subwavelength annular aperture arrays in metallic films. *Opt. Commun.*, 209:17–22, 2002.

[21] Y. Poujet, J. Salvi, and F. I. Baida. 90% extraordinary optical transmission in the visible range through annular aperture metallic arrays. *Opt. Lett.*, 32(20):2942–2944, 2007.

[22] F. I. Baida. Enhanced transmission through subwavelength metallic coaxial apertures by excitation of the tem mode. *Applied Phys. B*, 89(2-3):145–149, 2007. Rapid Communication.

[23] F. I. Baida, A. Belkhir, O. Arar, E. H. Barakat, J. Dahdah, C. Chemrouk, D. Van Labeke, C. Diebold, N. Perry, and M.-P. Bernal. Enhanced optical transmission by light coaxing: Mechanism of the tem-mode excitation. *Micron*, 41:742–745, 2010.

[24] W.-J. Tsay and D. M. Pozar. Application of the fdtd technique to periodic problems in scattering and radiation. *IEEE Microwave and Guided Wave Letters*, 3:250–252, 1993.

[25] A. Alexanian, N. J. Kolias, R. C. Compton, and R. A. York. Three-dimensional fdtd analysis of quasi-optical arrays using floquet boundary conditions and berenger's pml. *IEEE Microwave and Guided Wave Letters*, 6:138–140, 1996.

[26] F. I. Baida and D. Van Labeke. Three-dimensional structures for enhanced transmission through a metallic film: Annular aperture arrays. *Phys. Rev. B*, 67:155314, 2003.

[27] M. E. Veysoglu, R. T. Shin, and J. A. Kong. A finite-difference time-domain analysis of wave scattering from periodic surfaces: Oblique-incidence case. *J.Elect.Waves Appl.*, **7**:1595–607, 1993.

[28] J. A. Roden, S. D. Gedney, M. P. Kesler, J. G. Maloney, and P. H. Harms. Time-domain analysis of periodic structures at oblique incidence: orthogonal and nonorthogonal fdtd implementations. *IEEE Transactions on Microwave Theory and Techniques*, 46:420–427, 1998.

[29] Y.-C. A. Kao and R. G. Atkins. A finite difference-time domain approach for frequency selective surfaces at oblique incidence. *IEEE Antennas and Propagation Society International Symposium*, 2:1432–1435, 1996.

[30] Y.-C. A. Kao. *Finite-difference time domain modeling of oblique incidence scattering from periodic surfaces*. Master's thesis, Massachusetts Institute of Technology, 1997.

[31] J. A. Roden. *Electromagnetic analysis of complex structures using the fdtd technique in general curvilinear coordinates*. Ph.d. thesis, University of Kentucky, Lexington, KY, 1997.

[32] P. H. Harms, J. A. Roden, J. G. Maloney, and M. P. Kesler. Numerical analysis of periodic structures using the split-field algorithm. In *13th Annual Review of Progress in Applied Computational Electromagnetics*, pages 104–111, 1997.

[33] A. Aminian and Y. Rahmat-Samii. Spectral fdtd: a novel technique for the analysis of oblique incident plane wave on periodic structures. *IEEE Transactions on Antennas and Propagation*, 54:1818–1825, 2006.

[34] A. Belkhir and F. I. Baida. Three-dimensional finite-difference time-domain algorithm for oblique incidence with adaptation of perfectly matched layers and nonuniform meshing: Application to the study of a radar dome. *Phys. Rev. E*, 77(5):056701, 2008.

[35] M. Hamidi, F. I. Baida, A. Belkhir, and O. Lamrous. Implementation of the critical points model in a sfm-fdtd code working in oblique incidence. *J. Phys. D: Appl. Phys.*, 44(24):245101, 2011.

[36] F. I. Baida and A. Belkhir. Split-field fdtd method for oblique incidence study of periodic dispersive metallic structures. *Opt. Lett.*, 34(16):2453–2455, 2009.

[37] A. Belkhir, O. Arar, S. S. Benabbes, O. Lamrous, and F. I. Baida. Implementation of dispersion models in the split-field–finite-difference-time-domain algorithm for the study of metallic periodic structures at oblique incidence. *Phys. Rev. E*, 81(4):046705, 2010.

[38] J. Rybczynski, K. Kempa, A. Herczynski, Y. Wang, M. J. Naughton, Z. F. Ren, Z. P. Huang, D. Cai, and M. Giersig. Subwavelength waveguide for visible light. *Appl. Phys. Lett.*, 90:021104, 2007.

[39] T. Thio. Photonic devices: Coaxing light into small spaces. *Nature Nanotechnology*, 2:136–138, 2007.

[40] D. Van Labeke, D. Gérard, B. Guizal,F. I. Baida and L. Li. An angle-independent Frequency Selective Surface in the optical range *Opt. Express*, 14, 11945–11951, 2006.

[41] Samuel Nosal. *Modélisation de structures périodiques et matériaux artificiels. Application á la conception d'un radôme passe-bande*. PhD thesis, Ecole Centrale Paris, 2009.

[42] K. Kempa, X. Wang, Z. F. Ren, and M. J. Naughton. Discretely guided electromagnetic effective medium. *Appl. Phys. Lett.*, 92:043114, 2008.

Chapter 10:

Exact Modal Methods

Boris Gralak

Table of Contents:

Chapter 10

Exact Modal Methods

Boris Gralak

CNRS, Aix-Marseille Université, École Centrale Marseille, Institut Fresnel,
13397 Marseille Cedex 20, France
boris.gralak@fresnel.fr

10.1 Introduction

Exact modal method (EMM) has been proposed to take advantage of geometry of lamellar gratings. These gratings are made of rectangular rods periodically spaced which can be considered locally as periodic multilayered stacks (see figure 10.1). This simple geometry makes it possible to expand the electromagnetic field on the basis of "exact modes", and to obtain an exact representation of the permittivity. In this particular case, EMM can be more efficient than similar methods based on Fourier expansion (coupled-wave method [1] or Fourier modal method [2, 3], see also chapter 13) which may lead to poor convergence due to the discontinuous nature of both electromagnetic field and permittivity. This advantage of EMM becomes more important when the permittivity contrast is high, e.g. for metallic lamellar gratings, and for three-dimensional structures like woodpile photonic crystals [4].

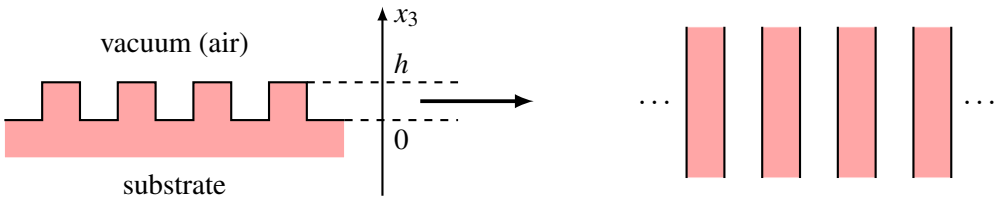


Figure 10.1: A lamellar grating made of a single lamellar layer on a substrate. The region corresponding to the lamellar layer, between the planes $x_3 = 0$ and $x_3 = h$, can be considered as the multilayered stack on the left.

Exact modal method has been introduced in 1981 in order to solve Maxwell's equations in presence of lamellar gratings made of dielectrics [5] and metals [6, 7, 8]. Since these pioneering works, a major contribution to this method is certainly its rigorous extension to conical mountings [9], on which is based an EMM for three-dimensional woodpile structures [4]. Another major development is the introduction of perfectly matched layers in order to model aperiodic systems met in integrated optics [10] (information can be found on the website of CAMFR).

In this chapter, a rigorous formulation of the exact modal method for lamellar structures is presented. In section 10.3, a special attention is paid to the continuation of the electromagnetic field inside a lamellar layer. In combination with the boundary conditions, this continuation provides a large class of solutions of Maxwell's equations in presence of lamellar gratings. In section 10.4, it is shown that, in each lamellar layer, there is a decoupling of the vector field equations into two independent scalar equations, which correspond to the ones of a multilayered stack (see figure 10.1). Numerical stacking algorithms are presented in section 10.5 and a numerical illustration of the EMM efficiency is proposed in section 10.6. The extension of the EMM to the particular case of lamellar gratings with infinitely conducting metal is presented in section 10.7. Finally, the techniques used for the calculation of the exact modes and the associated exact eigenvalues are reported in the appendix (section 10.8).

10.2 Notations

Throughout this chapter an orthonormal basis $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ is used: every vector $\boldsymbol{x}$ in $\mathbb{R}^3$ is described by its three components x_1, x_2 and x_3. It is shown how to obtain in the presence of a stack of lamellar layers, a large class of solutions $\boldsymbol{E}$ of the Helmholtz equation

$$\left[\, \omega^2 - \varepsilon^{-1} \boldsymbol{\nabla} \times \mu^{-1} \boldsymbol{\nabla} \times \right] \boldsymbol{E} = \boldsymbol{0}\,, \tag{10.1}$$

where ε is the permittivity, μ is the permeability, ω is the frequency and $\boldsymbol{\nabla}\times$ is the curl operator. All the media considered in this chapter are isotropic, and thus the permittivity and permeability reduce to scalar functions. The considered structure is independent of the variable x_2, and x_1-periodic with spatial period $\boldsymbol{d} = d\boldsymbol{e}_1$:

$$\varepsilon(\boldsymbol{x}+\boldsymbol{d}) = \varepsilon(\boldsymbol{x}) = \varepsilon(x_1, x_3)\,, \quad \mu(\boldsymbol{x}+\boldsymbol{d}) = \mu(\boldsymbol{x}) = \mu(x_1, x_3)\,, \quad \boldsymbol{x} \in \mathbb{R}^3\,. \tag{10.2}$$

The unit cell associated with this grating is $[0, d]$ and the one-dimensional lattice is $\left\{ n\mathbf{d} \,\middle|\, n \in \mathrm{Z} \right\}$. Then, a *lamellar grating is a stack in the direction x_3 of lamellar layers where ε and μ are both functions of the single variable x_1* (figure 10.2). In practice, each lamellar layer is made of infinite parallel rods with rectangular cross section (figure 10.2): the functions ε and μ are piecewise constant of the solely variable x_1.

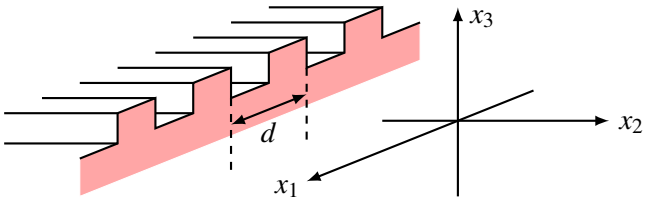


Figure 10.2: A lamellar grating made of a single lamellar layer on a substrate.

In order to obtain a set of first order differential equations from (10.1) a second field is defined:

$$\boldsymbol{H} = (\omega\mu)^{-1} \boldsymbol{\nabla} \times \boldsymbol{E}\,. \tag{10.3}$$

Note that this quantity differs from the usual "harmonic $\boldsymbol{H}$ field" by the complex number i. Solutions $\boldsymbol{E}$, $\boldsymbol{H}$ are investigated in the space of fields whose restrictions in every horizontal

plane (normal to $\boldsymbol{e}_3$) are square integrable:

$$\int_{\mathbb{R}^2} \left|\boldsymbol{F}(x_1,x_2,x_3)\right|^2 dx_1 dx_2 < \infty, \qquad x_3 \in \mathbb{R}, \tag{10.4}$$

where $\boldsymbol{F} = \boldsymbol{E}, \boldsymbol{H}$.

The first consequence of (10.4) is the possibility to perform a decomposition of the problem to take advantage of the spatial invariances of the system: a Floquet-Bloch decomposition with respect to the variable x_1,

$$\boldsymbol{F} \longrightarrow \tilde{\boldsymbol{F}}(k_1,x_1,x_2,x_3)\frac{1}{2\pi}\sum_{n\in Z} \exp[-ik_1 nd]\boldsymbol{F}(x_1+pd,x_2,x_3)\,, \tag{10.5}$$

where k_1 is the Bloch wave vector in the first Brillouin zone $[-\pi/d, \pi/d]$, and a Fourier decomposition with respect to the variable x_2,

$$\tilde{\boldsymbol{F}} \longrightarrow \widehat{\boldsymbol{F}}(k_1,x_1,k_2,x_3) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp[-ik_2x_2]\tilde{\boldsymbol{F}}(k_1,x_1,x_2,x_3)\,dx_2\,. \tag{10.6}$$

Thus solutions $\widehat{\boldsymbol{E}}$, $\widehat{\boldsymbol{H}}$ satisfy

$$\int_{[-\pi/d,\pi/d]} \left|\widehat{\boldsymbol{F}}(k_1,x_1,k_2,x_3)\right|^2 dx_1 < \infty, \qquad x_1,k_2,x_3 \in \mathbb{R}, \tag{10.7}$$

with the partial Bloch boundary condition

$$\widehat{\boldsymbol{F}}(k_1,x_1+d,k_2,x_3) = \exp[ik_1d]\widehat{\boldsymbol{F}}(k_1,x_1,k_2,x_3)\,, \tag{10.8}$$

where k_1 is fixed in $[-\pi/d, \pi/d]$.

The second consequence of (10.4) [or (10.7)] is that the restrictions to every horizontal plane of $\boldsymbol{\nabla}\times\boldsymbol{E}$ and $\boldsymbol{\nabla}\times\boldsymbol{H}$ are also locally square integrable [from (10.1,10.3)]. Then, for all $i,j = 1,2,3$ and $i \neq j$, E_i and H_i are continuous functions of the variable x_j. In particular, the tangential components E_1, E_2, H_1 and H_2 of $\boldsymbol{E}$ and $\boldsymbol{H}$ are continuous functions of the variable x_3. It follows that it is possible to solve Maxwell's equations in a stack of layers by the two following steps: the first step consists in solving Maxwell's equations in each layer independently and then the second step consists in connecting each independent solution using the continuity of E_1, E_2, H_1 and H_2.

With the definition (10.3), equation (10.1) is equivalent to the set of first order equations

$$\boldsymbol{E} = (\omega\varepsilon)^{-1}\boldsymbol{\nabla}\times\boldsymbol{H}\,, \quad \boldsymbol{H} = (\omega\mu)^{-1}\boldsymbol{\nabla}\times\boldsymbol{E}\,. \tag{10.9}$$

Let the 2×2 matrix σ and the two-components vector F_j defined by

$$\sigma = \omega\begin{bmatrix} 0 & \mu \\ \varepsilon & 0 \end{bmatrix}, \quad F_j = \begin{bmatrix} \tilde{E}_j \\ \tilde{H}_j \end{bmatrix}, \qquad j = 1,2,3\,. \tag{10.10}$$

Then, the first order equations (10.9) can be developed as

$$\begin{aligned} F_1 &= \sigma^{-1}\left[\partial_2 F_3 - \partial_3 F_2\right], \\ F_2 &= \sigma^{-1}\left[\partial_3 F_1 - \partial_1 F_3\right], \\ F_3 &= \sigma^{-1}\left[\partial_1 F_2 - \partial_2 F_1\right], \end{aligned} \tag{10.11}$$

where ∂_j is the partial derivative with respect to the variable x_j ($j = 1,2,3$). This last set of equations is exactly the same as (10.9) and, with some abuse of notations, it can written in the compact way $\boldsymbol{F} = \sigma^{-1}\boldsymbol{\nabla}\times\boldsymbol{F}$, with $\boldsymbol{F} = (F_1,F_2,F_3)$.

10.3 Continuation of the electromagnetic field

In this section two different formulations are presented to solve the equation $\boldsymbol{F} = \sigma^{-1}\boldsymbol{\nabla}\times\boldsymbol{F}$ in a lamellar layer located between the planes $x_3 = 0$ and $x_3 = h$. In practice, this solution is expressed as a relationship between $\boldsymbol{F}(0)$ and $\boldsymbol{F}(h)$. Note that the formulations presented in this section remain valid in the general case of cross gratings with two-dimensional periodicity [2].

10.3.1 Direct formulation: the transfer matrix

The starting point is the set of equations (10.11). Eliminating the components F_3, one obtains

$$\partial_3 F = iM\,F\,, \qquad F = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}, \qquad M = -i\begin{bmatrix} -\partial_1\sigma^{-1}\partial_2 & \sigma+\partial_1\sigma^{-1}\partial_1 \\ -\sigma-\partial_2\sigma^{-1}\partial_2 & \partial_2\sigma^{-1}\partial_1 \end{bmatrix}. \tag{10.12}$$

For a lamellar layer located between the planes $x_3 = 0$ and $x_3 = h$ (see figure 10.1), the functions ε and μ are x_3-independent for x_3 in $[0,h]$: then the matrix σ and the operator-valued matrix M are also x_3-independant for x_3 in $[0,h]$. Let L be the x_3-independant operator valued matrix which coincides with M in this single layer:

$$L = M(x_3)\,, \qquad x_3 \in [0,h]\,. \tag{10.13}$$

In a first step, it is assumed that (see the end of section 10.4 for a justification) the matrix L has a diagonal form and can be written as

$$L = V\,\lambda\,V^{-1}\,, \tag{10.14}$$

where the matrix V contains the eigenvectors $V_{\pm,n}$ of L, and λ is the diagonal matrix made of the associated eigenvalues $\lambda_{\pm,n}$:

$$LV_{\pm,n} = \lambda_{\pm,n}V_{\pm,n}\,. \tag{10.15}$$

Note that it is assumed that the matrix V is invertible: this is true in general for magnetodielectrics materials but, for example, this is not any more correct in the case where the system includes infinitely conduting metal. The sets of eigenvectors and eigenvalues are split into two parts according to the sign of the imaginary part of $\lambda_{\pm,n}$: $\mathrm{Im}\lambda_{+,n} > 0$ and $\mathrm{Im}\lambda_{-,n} < 0$. Let λ_+ (respectively λ_-) be the diagonal matrices containing the eigenvalues $\lambda_{+,n} > 0$ of L (respectively $\lambda_{-,n} > 0$): then

$$\lambda = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix}, \qquad \mathrm{Im}\lambda_+ > 0\,, \qquad \mathrm{Im}\lambda_- < 0\,. \tag{10.16}$$

This last condition on the imaginary part of eigenvalues is always realized if there is some absorption, *i.e.* $\mathrm{Im}\sigma > 0$, or if a small positive imaginary part is added to the frequency ω (in the later case the limit $\omega = \lim_{\eta\downarrow 0}(\omega + i\eta)$ is considered [11, 12]). The combination of (10.12) and (10.13) leads to the equation

$$\partial_3 F(x_3) = iLF(x_3)\,, \qquad x_3 \in [0,h]\,, \tag{10.17}$$

where the dependence on other variables has been omitted. Since L is x_3-independent, the "formal" solution of this equation is just

$$F(x_3) = \exp[iLx_3]F(0)\,, \qquad x_3 \in [0,h]\,. \tag{10.18}$$

This solution is denominated by "formal" since, at this stage, it is still necessary to check if it exists. Using the diagonal form (10.14) of the operator L, the expression (10.18) becomes

$$F(x_3) = V \exp[i\lambda x_3] V^{-1} F(0)\,. \tag{10.19}$$

Actually, the diagonal matrix $\exp[i\lambda x_3]$ is made of the two parts $\exp[i\lambda_{\pm} x_3]$ which have different behaviour. From (10.16), the part $\exp[i\lambda_+ x_3]$ is bounded by $\exp[-\mathrm{Im}\lambda_+ x_3] < 1$. On the contrary, the part $\exp[i\lambda_- x_3]$ is not bounded and, in general, the corresponding coefficients are growing towards infinity like exponential functions. Consequently, the transfer matrix $T(x_3)$ defined by

$$F(x_3) = T(x_3) F(0)\,, \qquad T(x_3) = V \exp[i\lambda x_3] V^{-1}\,, \tag{10.20}$$

has "infinite" coefficients and thus expressions like (10.18,10.19,10.20) have to be considered as purely "formal" and have to be handled cautiously. Numerically, the transfer matrix is truncated and, because its coefficients tend to infinity like exponential functions, it presents numerical instabilities which makes it difficult to use it. A numerical solution has been found to solve this problem with the definition of the S- and R-algorithms [13] (see section 10.5 for the numerical solution in the present case).

Finally, notice that the transfer matrix is occasionally derived from the matrix L^2 instead of L. Indeed, the assumption (10.14) on the diagonal form of L might be too strong and not rigorously true. According to notations (10.14), we denote by V and λ^2 the matrices containing the eigenvectors and eigenvalues of L^2:

$$L^2 = V \lambda^2 V^{-1}\,. \tag{10.21}$$

In that case, it used that equations (10.12) and (10.13) imply

$$\partial_3^2 F = -L^2 F\,, \qquad x_3 \in [0,h]\,. \tag{10.22}$$

The combination of the two last equations leads to

$$F(x_3) = V \cos[\lambda x_3] V^{-1} F(0) + V \lambda^{-1} \sin[\lambda x_3] V^{-1} (\partial_3 F)(0)\,. \tag{10.23}$$

Replacing $(\partial_3 F)(0)$ by $iLF(0)$, one obtains for the transfer matrix the following "formal" expression

$$T(x_3) = V \cos[\lambda x_3] V^{-1} + iV \lambda^{-1} \sin[\lambda x_3] V^{-1} L\,. \tag{10.24}$$

This equation is not "formally" equivalent to the first expression (10.20) derived from L. This equivalence requires the assumption (10.14) to become true, so that L can be replaced by $V\lambda V^{-1}$ above (and next the formal identity $\cos[\lambda x_3] + i \sin[\lambda x_3] = \exp[i\lambda x_3]$ has to be used).

10.3.2 Rigorous derivation of the continuation procedure

A rigorous formulation is based on the use of the Fourier transform with respect to the variable x_3 defined by

$$\mathscr{F}[F](k_3) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp[-ik_3 x_3] F(x_3)\, dx_3\,. \tag{10.25}$$

The function F is then deduced from its Fourier transform $\mathscr{F}[F]$ by

$$F(x_3) = \int_{\mathbb{R}} \exp[ik_3 x_3] \mathscr{F}[F](k_3)\, dk_3\,. \tag{10.26}$$

It is not suitable to perform directly the Fourier transform of the equation (10.12) since the matrix σ (and then M) is not independent of x_3 in $\mathbb{R}$. However, if equation (10.12) is multiplied by the characteristic function Ψ of the lamellar layer [so $\Psi(x_3) = 1$ for x_3 in $[0,h]$ and vanishes otherwise], then

$$\Psi(x_3)\partial_3 F(x_3) = \Psi(x_3) iM F(x_3) = \Psi(x_3) iL F(x_3) = iL\Psi(x_3)F(x_3)\,. \tag{10.27}$$

After this multiplication, a partial differential equation with the x_3-independent matrix L is obtained. The Fourier transform (10.25) of $\Psi\partial_3 F$ is

$$\begin{aligned}\mathscr{F}[\Psi\partial_3 F](k_3) &= \frac{1}{2\pi}\int_{\mathbb{R}} \exp[-ik_3x_3]\Psi(x_3)\partial_3 F(x_3)\,dx_3 \\ &= \frac{1}{2\pi}\int_0^h \exp[-ik_3x_3]\partial_3 F(x_3)\,dx_3 \\ &= \frac{1}{2\pi}\left\{\exp[-ik_3h]F(h) - F(0)\right\} + ik_3\mathscr{F}[\Psi F](k_3)\,,\end{aligned} \tag{10.28}$$

where the last line comes from an integration by parts. After this Fourier transform, equation (10.27) becomes

$$\frac{1}{2\pi}\left\{\exp[-ik_3h]F(h) - F(0)\right\} + ik_3\mathscr{F}[\Psi F](k_3) = iL\mathscr{F}[\Psi F](k_3) \tag{10.29}$$

or

$$\big[k_3 - L\big]\mathscr{F}[\Psi F](k_3) = \frac{1}{2i\pi}\left\{F(0) - \exp[-ik_3h]F(h)\right\}. \tag{10.30}$$

The operator $\big[k_3 - L\big]$ is always invertible if there is some absorption, *i.e.* $\mathrm{Im}\sigma > 0$, or if the limit $\omega = \lim_{\eta\downarrow 0}(\omega + i\eta)$ is considered (see [11, 12], this is equivalent to the property (10.16) on the eigenvalues λ since k_3 is purely real). Hence it is possible to write

$$\mathscr{F}[\Psi F](k_3) = \frac{1}{2i\pi}\frac{1}{k_3 - L}\left\{F(0) - \exp[-ik_3h]F(h)\right\}, \tag{10.31}$$

The final step is to apply the inverse Fourier transform (10.26): for x_3 in $[0,h]$,

$$\begin{aligned}\Psi(x_3)F(x_3) = &\ \frac{1}{2i\pi}\left[\int_{\mathbb{R}} \exp[ik_3x_3]\,\frac{1}{k_3 - L}\,dk_3\right] F(0) \\ - &\ \frac{1}{2i\pi}\left[\int_{\mathbb{R}} \exp[ik_3(x_3 - h)]\,\frac{1}{k_3 - L}\,dk_3\right] F(h)\,.\end{aligned} \tag{10.32}$$

Again, it is assumed that the operator L can be written $L = V\lambda V^{-1}$ (10.14). Replacing the matrix L by its diagonal form, the last expression becomes

$$\begin{aligned}\Psi(x_3)F(x_3) = &\ \frac{1}{2i\pi}V\left[\int_{\mathbb{R}} \exp[ik_3x_3]\,\frac{1}{k_3 - \lambda}\,dk_3\right] V^{-1}F(0) \\ - &\ \frac{1}{2i\pi}V\left[\int_{\mathbb{R}} \exp[ik_3(x_3 - h)]\,\frac{1}{k_3 - \lambda}\,dk_3\right] V^{-1}F(h)\,.\end{aligned} \tag{10.33}$$

The integrations above are performed by adding to the real axis of k_3 a semi-circle with infinite radius (in the complex plane of k_3) on which the integrals vanish. For the first term with $F(0)$,

the complex number k_3 must have positive imaginary part (x_3 is positive), so that the real axis is closed by a semi-circle in the upper half plane (see the red path on figure 10.3). In this case, the solely eigenvalues contained in λ_+ generate contributions in the integral. For the second term with $F(h)$, the complex number k_3 must have negative imaginary part ($x_3 - h$ is positive), so that the real axis is closed by a semi-circle in the lower half plane (see the blue path on figure 10.3). Here, the integral is given by the eigenvalues contained in λ_-. Let $P_\pm$ be the projectors

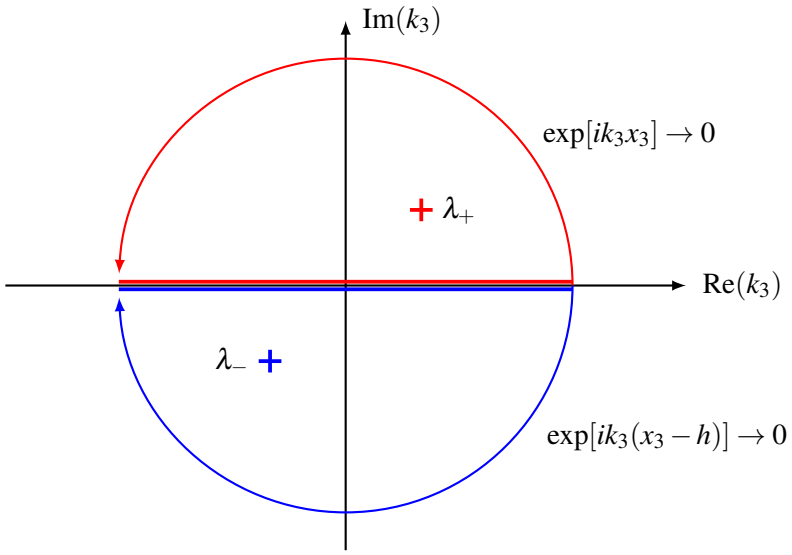


Figure 10.3: Integration in the complex plane of k_3.

upon the spaces corresponding respectively to eigenvalues $\lambda_\pm$:

$$P_+\lambda = \begin{bmatrix} \lambda_+ & 0 \\ 0 & 0 \end{bmatrix}, \qquad P_-\lambda = \begin{bmatrix} 0 & 0 \\ 0 & \lambda_- \end{bmatrix}. \tag{10.34}$$

Then, after the integration over k_3, expression (10.33) yields

$$\Psi(x_3)F(x_3) = V\,P_+ \exp[i\lambda_+ x_3]\,V^{-1}F(0) + V\,P_- \exp[i\lambda_-(x_3-h)]\,V^{-1}F(h)\,. \tag{10.35}$$

This expression is always well defined since the integration in the complex plane of k_3 imposes that all the complex exponential functions decrease:

$$\left\| \exp[i\lambda_+ x_3] \right\| \leq 1\,, \qquad \left\| \exp[i\lambda_-(x_3-h)] \right\| \leq 1\,. \tag{10.36}$$

Considering the rigorous expression (10.35) at $x_3 = 0$ and $x_3 = h$ and using that $P_- + P_+$ is the identity, one obtains

$$\begin{aligned} V\,P_-\,V^{-1}F(0) &= V\,P_- \exp[-i\lambda_- h]\,V^{-1}F(h)\,, \\ V\,P_+\,V^{-1}F(h) &= V\,P_+ \exp[i\lambda_+ h]\,V^{-1}F(0)\,. \end{aligned} \tag{10.37}$$

These two relationships provides a rigorous way to deduce $F(0)$ from $F(h)$ and conversely.

As in the previous section, the continuation procedure is also derived from the diagonal form (10.21) of L^2. Equation (10.22) is multiplied by $\Psi(x_3)$ to provide an expression similar to (10.27)

$$\Psi(x_3)\partial_3 F(x_3) = -L^2\,\Psi(x_3)F(x_3)\,. \tag{10.38}$$

Then, the Fourier transform (10.25) is applied to this equation. Using that

$$\begin{aligned}\mathscr{F}[\Psi\partial_3^2 F](k_3) &= \frac{1}{2\pi}\int_{\mathbb{R}} \exp[-ik_3x_3]\Psi(x_3)\partial_3^2 F(x_3)\,dx_3 \\ &= -k_3^2\mathscr{F}[\Psi F](k_3) + ik_3\frac{1}{2\pi}\left\{\exp[-ik_3h]F(h) - F(0)\right\} \\ &+\frac{1}{2\pi}\left\{\exp[-ik_3h](\partial_3 F)(h) - (\partial_3 F)(0)\right\},\end{aligned} \tag{10.39}$$

and replacing $(\partial_3 F)(x_3)$ by $iLF(x_3)$, equation (10.38) implies

$$\begin{aligned}\mathscr{F}[\Psi F](k_3) &= \frac{1}{2i\pi}\frac{1}{k_3^2-L^2}\,k_3\left\{F(0) - \exp[-ik_3h]F(h)\right\} \\ &= \frac{1}{2i\pi}\frac{1}{k_3^2-L^2}\left\{LF(0) - \exp[-ik_3h]LF(h)\right\}.\end{aligned} \tag{10.40}$$

Next, the inverse Fourier transform (10.26) is performed for x_3 in $[0,h]$ and the diagonal form (10.21) is used:

$$\begin{aligned}\Psi(x_3)F(x_3) = &\frac{1}{2i\pi}V\int_{\mathbb{R}}\frac{k_3}{k_3^2-\lambda^2}\left\{\exp[ik_3x_3]V^{-1}F(0) - \exp[ik_3(x_3-h)]V^{-1}F(h)\right\}dx_3 \\ &+ \frac{1}{2i\pi}V\int_{\mathbb{R}}\frac{1}{k_3^2-\lambda^2}\left\{\exp[ik_3x_3]V^{-1}LF(0) - \exp[ik_3(x_3-h)]V^{-1}LF(h)\right\}dx_3\,.\end{aligned} \tag{10.41}$$

Again, the integrations above are calculated by adding to the real axis of k_3 a semi-circle with infinite radius (in the complex plane of k_3) on which the integrals vanish. Without loss of generality, it is considered that the square root of the eigenvalues in λ^2 have non-zero imaginary part: let $\sqrt{\lambda^2}$ be the square root of λ^2 with positive imaginary part. For the terms with $F(0)$, the real axis is closed by a semi-circle in the upper half plane, and the eigenvalues with positive imaginary part $\sqrt{\lambda^2}$ lead to contributions in the integrals. For the terms with $F(h)$, the real axis is closed by a semi-circle in the lower half plane, and the integrals are given by the eigenvalues with negative imaginary part, *i.e.* $-\sqrt{\lambda^2}$. Calculations of integrals over k_3 lead to

$$\begin{aligned}\Psi(x_3)F(x_3) = &\;V\frac{1}{2}\exp[i\sqrt{\lambda^2}x_3]V^{-1}F(0) + V\frac{1}{2}\exp[-i\sqrt{\lambda^2}(x_3-h)]V^{-1}F(h) \\ &+ V\frac{1}{2\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}x_3]V^{-1}LF(0) - V\frac{1}{2\sqrt{\lambda^2}}\exp[-i\sqrt{\lambda^2}(x_3-h)]V^{-1}LF(h)\,.\end{aligned} \tag{10.42}$$

This equation is evaluated at $x_3 = 0$ and $x_3 = h$:

$$\begin{aligned}F(0) = &\;V\exp[i\sqrt{\lambda^2}h]V^{-1}F(h) + V\frac{1}{\sqrt{\lambda^2}}V^{-1}LF(0) - V\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1}LF(h)\,, \\ F(h) = &\;V\exp[i\sqrt{\lambda^2}h]V^{-1}F(0) + V\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1}LF(0) - V\frac{1}{\sqrt{\lambda^2}}V^{-1}LF(h)\,.\end{aligned} \tag{10.43}$$

Thanks to the technique based on the Fourier transform, all the exponential functions in these expressions must be well-defined. Indeed, the imaginary part of $\sqrt{\lambda^2}$ is positive and all the exponential functions decrease. Equations (10.43) will be used to construct a stable numerical algorithm to stack several lamellar layers.

10.4 Exact eigenmodes and eigenvalues method

The different solutions (10.20), (10.24, (10.37) and (10.43), established in the previous section, are provided from the knowledge of the sets of eigenmodes and eigenvalues of the operator L (or L^2). In this section, it is shown how these eigenmodes $V_{\pm,n}$ and eigenvalues $\lambda_{\pm,n}$ of L can be exactly determined in a lamellar layer located between the planes $x = 0$ and $x_3 = h$. The starting point is equation (10.17):

$$\partial_3 F = iLF\,, \qquad L = M(x_3)\,, \qquad x_3 \in [0,h]. \tag{10.44}$$

Let ε_1, μ_1 and σ_1 be the functions which coincide with respectively ε, μ and σ in the considered lamellar layer for x_3 in $[0,h]$. In the considered lamellar layer, they are functions of the solely variable x_1 (see figure 10.2):

$$\varepsilon_1(x_1) = \varepsilon(x_1,x_3)\,, \quad \mu_1(x_1) = \mu(x_1,x_3)\,, \quad \sigma_1(x_1) = \sigma(x_1,x_3)\,, \qquad x_3 \in [0,h]. \tag{10.45}$$

According to expression (10.12), the operator L is now

$$L = -i \begin{bmatrix} -\partial_1\sigma_1^{-1}\partial_2 & \sigma_1 + \partial_1\sigma_1^{-1}\partial_1 \\ -\sigma_1 - \partial_2\sigma_1^{-1}\partial_2 & \partial_2\sigma_1^{-1}\partial_1 \end{bmatrix}, \tag{10.46}$$

and its square is

$$L^2 = -\begin{bmatrix} -\sigma_1^2 - \partial_1\sigma_1^{-1}\partial_1\sigma_1 - \sigma_1\partial_2\sigma_1^{-1}\partial_2 & \sigma_1\partial_2\sigma_1^{-1}\partial_1 - \partial_1\sigma_1^{-1}\partial_2\sigma_1 \\ \sigma_1\partial_1\sigma_1^{-1}\partial_2 - \partial_2\sigma_1^{-1}\partial_1\sigma_1 & -\sigma_1^2 - \partial_2\sigma_1^{-1}\partial_2\sigma_1 - \sigma_1\partial_1\sigma_1^{-1}\partial_1 \end{bmatrix}. \tag{10.47}$$

Since the matrix σ_1 is x_2-independent, the equality $\sigma_1\partial_2\sigma_1^{-1} = \partial_2 = \sigma_1^{-1}\partial_2\sigma_1$ holds, and the expression above becomes

$$L^2 = \begin{bmatrix} \sigma_1^2 + \partial_2^2 + \partial_1\sigma_1^{-1}\partial_1\sigma_1 & 0 \\ \partial_2\sigma_1^{-1}\partial_1\sigma_1 - \sigma_1\partial_1\sigma_1^{-1}\partial_2 & \sigma_1^2 + \partial_2^2 + \sigma_1\partial_1\sigma_1^{-1}\partial_1 \end{bmatrix}. \tag{10.48}$$

This expression shows that the components F_1 can be decoupled from the components F_2 in the lamellar layer. Indeed, it implies

$$\partial_3^2 F_1 = -KF_1\,, \qquad K = \sigma_1^2 + \partial_2^2 + \partial_1\sigma_1^{-1}\partial_1\sigma_1\,. \tag{10.49}$$

Moreover, each component of F_1, *i.e.* E_1 and H_1, can be also decoupled since the operator K is diagonal:

$$K = \begin{bmatrix} K_{\varepsilon_1} & 0 \\ 0 & K_{\mu_1} \end{bmatrix}, \tag{10.50}$$

where

$$\begin{aligned} K_{\varepsilon_1} &= \omega^2\varepsilon_1\mu_1 + \partial_2^2 + \partial_1\varepsilon_1^{-1}\partial_1\varepsilon_1\,, \\ K_{\mu_1} &= \omega^2\varepsilon_1\mu_1 + \partial_2^2 + \partial_1\mu_1^{-1}\partial_1\mu_1\,. \end{aligned} \tag{10.51}$$

Here, it is important to notice that the two operators K_{ε_1} and K_{μ_1} correspond to the ones of a one-dimensional multilayered stack for respectively p- and s-polarization. This makes it possible to calculate the exact eigenmodes and eigenvalues of K_{ε_1} and K_{μ_1} (see appendix) and thus the ones of K. Thus the continuation procedure presented in section 10.3.2 can be applied to

equation (10.49). It provides relationships between the fields $F_1(0)$, $F_1(h)$ and their derivative with respect to x_3, *i.e.* $[\partial_3 F_1](0)$ and $[\partial_3 F_1](h)$ (it is recalled that $\partial_3 F = iLF$ in section 10.3.2).

To complete the derivation of the method, it is necessary to express the component F_2 of the field from F_1 and $\partial_3 F_1$. A starting relationship is obtained from (10.44) and (10.46):

$$\partial_3 F_1 = -\partial_1 \sigma_1^{-1} \partial_2 F_1 + \left[\sigma_1 + \partial_1 \sigma_1^{-1} \partial_1\right] F_2 \,. \tag{10.52}$$

This equation is equivalent to

$$\left[1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}\right] \sigma_1 F_2 = \partial_3 F_1 + \partial_1 \sigma_1^{-1} \partial_2 F_1 \,. \tag{10.53}$$

Here, it is remarked that the operator $[1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}]$ is invertible since it equals $[K - \partial_2^2]\sigma_1^{-2}$ where σ_1 is invertible as well as $[K - \partial_2^2]$ (the eigenvalues of K have non-zero imaginary part). Moreover, the operator $[1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}]$ commutes with ∂_3, ∂_2 and $\partial_1 \sigma_1^{-1}$: hence

$$\begin{aligned}
\sigma_1 F_2 &= \frac{1}{1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}} \partial_3 F_1 + \frac{1}{1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}} \partial_1 \sigma_1^{-1} \partial_2 F_1 \\
&= \frac{1}{1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}} \partial_3 F_1 + \partial_2 \partial_1 \sigma_1^{-1} \frac{1}{1 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1^{-1}} F_1 \\
&= \sigma_1^2 \frac{1}{\sigma_1^2 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1} \partial_3 F_1 + \partial_2 \partial_1 \sigma_1 \frac{1}{\sigma_1^2 + \partial_1 \sigma_1^{-1} \partial_1 \sigma_1} F_1 \,.
\end{aligned} \tag{10.54}$$

After the Fourier decomposition with respect to the variable x_2, the operator ∂_2 becomes ik_2. Then, the inverse operator above is expressed using the diagonal form $K = V\lambda^2 V^{-1}$, and the component F_2 becomes

$$F_2 = \sigma_1 V \frac{1}{\lambda^2 + k_2^2} V^{-1} \partial_3 F_1 + ik_2 \sigma_1^{-1} \partial_1 \sigma_1 V \frac{1}{\lambda^2 + k_2^2} V^{-1} F_1 \,. \tag{10.55}$$

Finally, it is stressed that the coefficients of the operators $\sigma_1 V$ and $\sigma_1^{-1} \partial_1 \sigma_1 V$ above can be calculated exactly from the knowledge of the exact eigenmodes. Indeed, the technique presented in appendix shows that the determination of the eigenmodes ϕ_p lies on the calculations of functions $\sigma_1 \phi_p$ and $\sigma_1^{-1} \partial_1 \sigma_1 \phi_p$. The final expression of F_2 in terms of F_1 and $\partial_3 F_1$ is then

$$F_2 = U \frac{1}{\lambda^2 + k_2^2} V^{-1} \partial_3 F_1 + ik_2 W \frac{1}{\lambda^2 + k_2^2} V^{-1} F_1 \,, \qquad U = \sigma_1 V \,, \qquad W = \sigma_1^{-1} \partial_1 \sigma_1 V \,. \tag{10.56}$$

To conclude this section, it has been shown that the field components F_1, $\partial_3 F_1$ and F_2 can be expressed exactly in a lamellar layer: F_1 and $\partial_3 F_1$ are provided by equations (10.42) and (10.43) (where $F \to F_1$, $LF \to -i\partial_3 F_1$, and $K = V\lambda^2 V^{-1}$); F_2 is given by equation (10.56). These expressions are based on the knowledge of the eigenvectors V of K, the eigenvalues λ^2 of K and the operators $U = \sigma_1 V$ and $W = \sigma_1^{-1} \partial_1 \sigma_1 V$. All these quantities can be calculated exactly, as shown in appendix.

Finally, it is stressed that the propagation constants inside the lamellar layer are the square roots of the eigenvalues of the operator K, *i.e.* $\pm\sqrt{\lambda^2}$. If there is some absorption, *i.e.* $\mathrm{Im}\sigma > 0$, or if a small positive imaginary part is added to the frequency ω (the limit $\omega = \lim_{\eta \downarrow 0}(\omega + i\eta)$ is considered [11, 12]), the eigenvalues λ^2 of K cannot be purely real. It follows that the propagation constants $\pm\sqrt{\lambda^2}$ have non zero imaginary parts and, moreover, the half of them ($+\sqrt{\lambda^2}$) have strictly positive imaginary part and the second half of them ($-\sqrt{\lambda^2}$) have strictly negative imaginary part. In this case, the assumption (10.16) on the eigenvalues of L is well justified.

10.5 Numerical algorithm

A general solution for numerical algorithm has been proposed by L. Li in the case of modal methods of gratings [3]. This solution is based on the definition of S or R matrices which are well-conditioned.

10.5.1 R matrix for a single lamellar layer

In this section, the expression of a R matrix associated with a lamellar layer is established: it is defined by the relationship

$$\begin{bmatrix} F_1(0) \\ F_1(h) \end{bmatrix} = R \begin{bmatrix} F_2(0) \\ F_2(h) \end{bmatrix}. \tag{10.57}$$

First, equation (10.43) is used to provide an expression of the solution of (10.49):

$$\begin{aligned} F_1(0) - V\exp[i\sqrt{\lambda^2}h]V^{-1}F_1(h) &= -iV\frac{1}{\sqrt{\lambda^2}}V^{-1}\partial_3F_1(0) + iV\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1}\partial_3F_1(h), \\ F_1(h) - V\exp[i\sqrt{\lambda^2}h]V^{-1}F_1(0) &= -iV\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1}\partial_3F_1(0) + iV\frac{1}{\sqrt{\lambda^2}}V^{-1}\partial_3F_1(h). \end{aligned} \tag{10.58}$$

This set of equations is written using 2×2 matrices:

$$A\begin{bmatrix} F_1(0) \\ F_1(h) \end{bmatrix} = B\begin{bmatrix} \partial_3F_1(0) \\ \partial_3F_1(h) \end{bmatrix}, \tag{10.59}$$

where

$$A = \begin{bmatrix} 1 & -V\exp[i\sqrt{\lambda^2}h]V^{-1} \\ -V\exp[i\sqrt{\lambda^2}h]V^{-1} & 1 \end{bmatrix} \tag{10.60}$$

and

$$B = \begin{bmatrix} -iV\frac{1}{\sqrt{\lambda^2}}V^{-1} & iV\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1} \\ -iV\frac{1}{\sqrt{\lambda^2}}\exp[i\sqrt{\lambda^2}h]V^{-1} & iV\frac{1}{\sqrt{\lambda^2}}V^{-1} \end{bmatrix}. \tag{10.61}$$

Next, from (10.56), the field ∂_3F_1 is related to F_1 and F_2 from

$$\partial_3F_1 = -ik_2V[\lambda^2+k_2^2]U^{-1}W\frac{1}{\lambda^2+k_2^2}V^{-1}F_1 + V[\lambda^2+k_2^2]U^{-1}F_2. \tag{10.62}$$

Defining the two matrices

$$C = \begin{bmatrix} -ik_2V[\lambda^2+k_2^2]U^{-1}W\frac{1}{\lambda^2+k_2^2}V^{-1} & 0 \\ 0 & -ik_2V[\lambda^2+k_2^2]U^{-1}W\frac{1}{\lambda^2+k_2^2}V^{-1} \end{bmatrix} \tag{10.63}$$

and

$$D = \begin{bmatrix} V[\lambda^2+k_2^2]U^{-1} & 0 \\ 0 & V[\lambda^2+k_2^2]U^{-1} \end{bmatrix}, \tag{10.64}$$

the relationship (10.59) becomes

$$A \begin{bmatrix} F_1(0) \\ F_1(h) \end{bmatrix} = BC \begin{bmatrix} F_1(0) \\ F_1(h) \end{bmatrix} + BD \begin{bmatrix} F_2(0) \\ F_2(h) \end{bmatrix}. \tag{10.65}$$

Finally, according to the definition (10.57), the R matrix is given by

$$R = \frac{1}{A-BC} BD. \tag{10.66}$$

It is stressed that the R matrix is numerically stable. Indeed, exponential functions always have arguments such that they decrease and, when inverted, they are always added to well-conditioned functions. For example, in the matrix $[A-BC]$, one can check that the diagonal blocs are well-conditioned since they do not contain any exponential function, while the off-diagonal blocs decrease exponentially. When inverted, this matrix will have the same behaviour with well-conditioned diagonal blocs and exponentially decreasing off-diagonal blocs.

10.5.2 R matrix for a stack of lamellar layers

Here, a system made of two lamellar layers is considered. The first layer is located between the planes $x_3 = -h_1$ and $x_3 = 0$, and the second layer between the planes $x_3 = 0$ and $x_3 = h_2$. Let R_1 and R_2 be the R matrices associated with these layers:

$$\begin{bmatrix} F_1(-h_1) \\ F_1(0) \end{bmatrix} = R_1 \begin{bmatrix} F_2(-h_1) \\ F_2(0) \end{bmatrix}, \qquad \begin{bmatrix} F_1(0) \\ F_1(h_2) \end{bmatrix} = R_2 \begin{bmatrix} F_2(0) \\ F_2(h_2) \end{bmatrix}. \tag{10.67}$$

Then, the R matrix associated with the stack of the two layers is determined by eliminating the components $F_1(0)$ and $F_2(0)$ in the equations above. Denoting by $R_{1,ij}$ and $R_{2,ij}$ $(i,j=1,2)$ the blocs of R_1 and R_2,

$$R_1 = \begin{bmatrix} R_{1,11} & R_{1,12} \\ R_{1,21} & R_{1,22} \end{bmatrix}, \qquad R_2 = \begin{bmatrix} R_{2,11} & R_{2,12} \\ R_{2,21} & R_{2,22} \end{bmatrix}, \tag{10.68}$$

the expression of R is given by

$$R = \begin{bmatrix} R_{1,11} - R_{1,12} \dfrac{1}{R_{1,22}-R_{2,11}} R_{1,21} & R_{1,12} \dfrac{1}{R_{1,22}-R_{2,11}} R_{2,12} \\ -R_{2,21} \dfrac{1}{R_{1,22}-R_{2,11}} R_{1,21} & R_{2,22} - R_{2,21} \dfrac{1}{R_{1,22}-R_{2,11}} R_{2,12} \end{bmatrix}. \tag{10.69}$$

Again, one can check that the algorithm is stable since the only inverted blocs are the diagonal ones, which are well-conditioned.

10.6 Numerical application

A simple numerical example is considered to put the exact modal method to the test. The structure is made of a set of rectangular rods with dielectric constant $\varepsilon_{1,1}/\varepsilon_0 = 12.96$ (corresponding to the index 3.6 for Si at optical wavelengths), width $w_{1,1} = 0.28d$ and height $h = d/(2\sqrt{2})$, where d is the spatial period of the grating (see figure 10.4). This lamellar grating

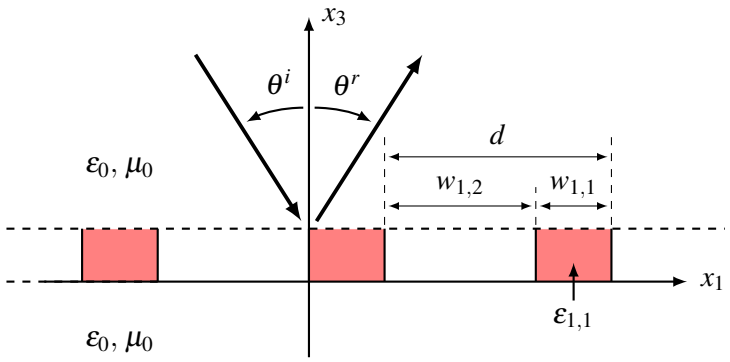


Figure 10.4: The considered structure for the numerical example: a single layer made of rectangular rods.

is illuminated by a plane wave with an incident angle $\theta^i = 45°$. The oscillating frequency is $\omega = 2\pi/(d\sqrt{\varepsilon_0\mu_0})$, which corresponds to a wavelength equal to the spatial period d. The efficiency diffracted in the order zero, *i.e.* at the reflected angle $\theta^r = \theta^i = 45°$ is calculated for both s and p-polarizations which correspond respectively to the electric and magnetic fields reduced to a single component along the invariance axis x_2. Each component of the electromagnetic field is described by a finite number $(2n+1)$ of exact modes. Reflected efficiency in the order zero for different values of the number of exact modes $(2n+1)$ is represented on figure 10.5. These curves show that these efficiencies differ from their converged value with less than one percent from $(2n+1) = 7$ in s-polarization and $(2n+1) = 5$ in p-polarization (the converged values are 0.7323 and 0.9487 in s and p polarizations). This convergence is found to be faster than in the case of the modal method with Fourier basis [2] (see also chapter 13 for the Fourier modal method) where an error smaller than one percent is obtained from $(2n+1) = 17$ and $(2n+1) = 27$ in s and p polarizations respectively (the method [2] contains all the techniques to improve the convergence of the truncated Fourier series [14]). This improvement of the convergence resulting from the use of the exact modes becomes a significant advantage when three dimensional woodpile structures are considered [4] (the total number of modes $(2n+1)^2$ can be reduced by a factor of 10).

10.7 Lamellar gratings including infinitely conducting metal

The motivations for studying lamellar metallic gratings are numerous. Indeed periodic metallic structures are good candidates for extraordinary transmission [15, 16], compact antennas [17], modified local density of states [18, 19, 20], negative index materials [21, 22], *et c.* However, the use of the EMM (as well as the Fourier modal method) leads to numerical instabilities, even if S or R algorithms [13] are implemented.

The method presented in previous sections is extended in order to obtain a suitable model for structures including infinitely conducting metal. Indeed, it necessary to modify the numerical algorithm to prevent the EMM from numerical instabilities. A convergence test shows that the method converges rapidly and is numerically stable. Finally, a comparison of a field map with the fictitious sources method shows perfect agreement.

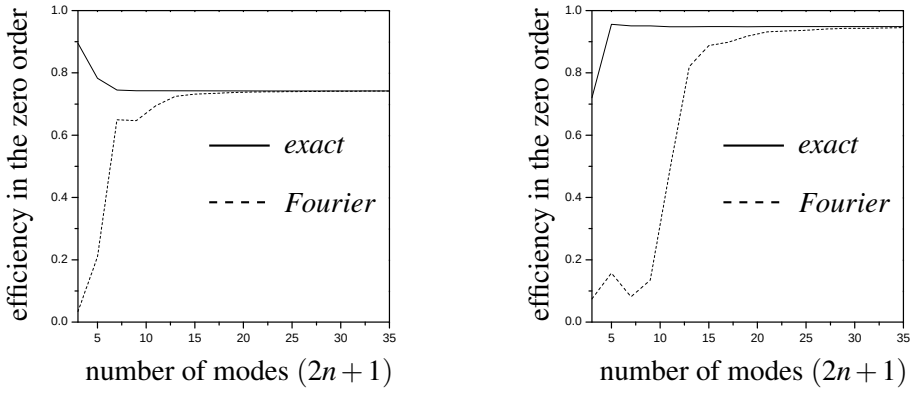


Figure 10.5: Efficiency in the zero order for s-polarization (left panel) and p-polarization (right panel).

10.7.1 Background

The set of first order equations (10.9) representing the time harmonic Maxwell's equations is considered in non magnetic media ($\mu = \mu_0$):

$$\boldsymbol{E} = (\omega\varepsilon)^{-1}\boldsymbol{\nabla}\times\boldsymbol{H}\,, \quad \boldsymbol{H} = (\omega\mu_0)^{-1}\boldsymbol{\nabla}\times\boldsymbol{E}\,. \tag{10.70}$$

The permittivity function ε is well-defined for linear (eventually dispersive and absorptive) dielectric materials and, in domains with infinitely conducting metal, the electric field must vanish. The two different regimes, dielectric and infinitely conducting, can be compiled by defining the characteristic function

$$\Psi_a = \begin{cases} 1 & \text{in dielectric materials}\,, \\ 0 & \text{in infinitely conducting metal.} \end{cases} \tag{10.71}$$

Thus, for systems including dielectrics and infinitely conducting metals, the time harmonic Maxwell's equations can be written

$$\boldsymbol{E} = \Psi_a(\omega\varepsilon)^{-1}\boldsymbol{\nabla}\times\boldsymbol{H}\,, \quad \boldsymbol{H} = (\omega\mu_0)^{-1}\boldsymbol{\nabla}\times\boldsymbol{E}\,. \tag{10.72}$$

With this set of equations, all the procedure derived in sections 10.3 and 10.4 remains valid provided ε and ε^{-1} are respectively replaced by $\Psi_a\varepsilon$ and $\Psi_a\varepsilon^{-1}$, and the matrix V^{-1} is replaced by $V^\dagger$, the adjoint of V. Indeed, from the expression of the exact eigenvalues and eigenvectors given in see appendix 10.8.5, the exact eigenvectors vanish in the interval $[a,d]$ and thus they cannot form a complete set of the Hilbert space of square integrable functions on the $[0,d]$. Consequently, it is not possible to develop all square integrable function on $[0,d]$ from the exact eigenvectors which span only the interval $[a,d]$, and thus the matrix (operator) V is not invertible.

As indicated in [23], this non-invertible matrix V is at the origin of numerical instabilities when several lamellar layers are stacked. A solution is to consider the "impedance" algorithm presented in [24] and [23].

10.7.2 Impedance algorithm

The impedance algorithm is based on the use of the "impedance" matrix which is defined by

$$\begin{bmatrix} E(0) \\ E(h) \end{bmatrix} = Z \begin{bmatrix} H(0) \\ H(h) \end{bmatrix}, \qquad E = \begin{bmatrix} \tilde{E}_1 \\ \tilde{E}_2 \end{bmatrix}, \qquad H = \begin{bmatrix} \tilde{H}_1 \\ \tilde{H}_2 \end{bmatrix}. \tag{10.73}$$

The expression of the impedance matrix Z can be determined from equation (10.65), or the expression

$$(A - BC) \begin{bmatrix} F_1(0) \\ F_1(h) \end{bmatrix} = BD \begin{bmatrix} F_2(0) \\ F_2(h) \end{bmatrix}, \tag{10.74}$$

where, in the expressions of A, B, C and D, the inverses V^{-1} and U^{-1} are replaced respectively by $V^{\dagger}$ and $U^{\dagger}\sigma_1^{-1}$.

Now, let Z_1 and Z_2 be the Z matrices associated with two layers located between the planes $x_3 = -h_1$ and $x_3 = 0$, and $x_3 = 0$ and $x_3 = h_2$:

$$\begin{bmatrix} E(-h_1) \\ E(0) \end{bmatrix} = Z_1 \begin{bmatrix} H(-h_1) \\ H(0) \end{bmatrix}, \qquad \begin{bmatrix} E(0) \\ E(h_2) \end{bmatrix} = Z_2 \begin{bmatrix} H(0) \\ H(h_2) \end{bmatrix}. \tag{10.75}$$

Then, the Z matrix associated with the stack of the two layers is determined by eliminating the components $E(0)$ and $H(0)$ in the equations above. Denoting by $Z_{1,ij}$ and $Z_{2,ij}$ $(i, j = 1, 2)$ the blocs of Z_1 and Z_2,

$$Z_1 = \begin{bmatrix} Z_{1,11} & Z_{1,12} \\ Z_{1,21} & Z_{1,22} \end{bmatrix}, \qquad Z_2 = \begin{bmatrix} Z_{2,11} & Z_{2,12} \\ Z_{2,21} & Z_{2,22} \end{bmatrix}, \tag{10.76}$$

the expression of Z is given by

$$Z = \begin{bmatrix} Z_{1,11} - Z_{1,12} \dfrac{1}{Z_{1,22} - Z_{2,11}} Z_{1,21} & Z_{1,12} \dfrac{1}{Z_{1,22} - Z_{2,11}} Z_{2,12} \\ -Z_{2,21} \dfrac{1}{Z_{1,22} - Z_{2,11}} Z_{1,21} & Z_{2,22} - Z_{2,21} \dfrac{1}{Z_{1,22} - Z_{2,11}} Z_{2,12} \end{bmatrix}. \tag{10.77}$$

More details about the method and the structures which can be modelled can be found in references [24] and [23].

10.7.3 Numerical example

The considered lamellar grating is a set of periodically-spaced and infinitely conducting "F" shaped scatterers, embedded in vacuum, see figure 10.6. This grating is illuminated by a plane

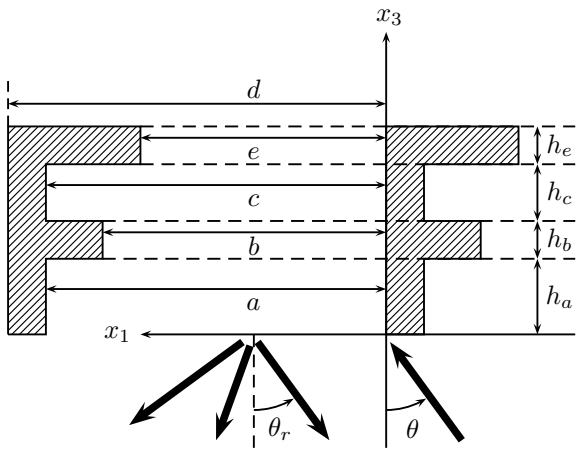


Figure 10.6: The considered lamellar grating: The spatial period is $d = 20.0u$, where u is an arbitrary unit; The four layers have air width $a = 18.0u$, $b = 15.0u$, $c = 18.0u$, $e = 13.0u$, and thickness $h_a = 4.0u$, $h_b = 2.0u$, $h_c = 3.0u$ and $h_e = 2.0u$.

wave with wavelength equal to $0.1d$, corresponding to the tenth of the spatial period. The incident angle of this plane wave is $\theta = 45$ degrees, and the conical angle $\varphi = 30$ degrees. Finally, the incident field is *s*-polarized, *i.e.* the electric field is perpendicular to the incident plane. To complete this first test, the total reflectivity has been represented on figure 10.7 as

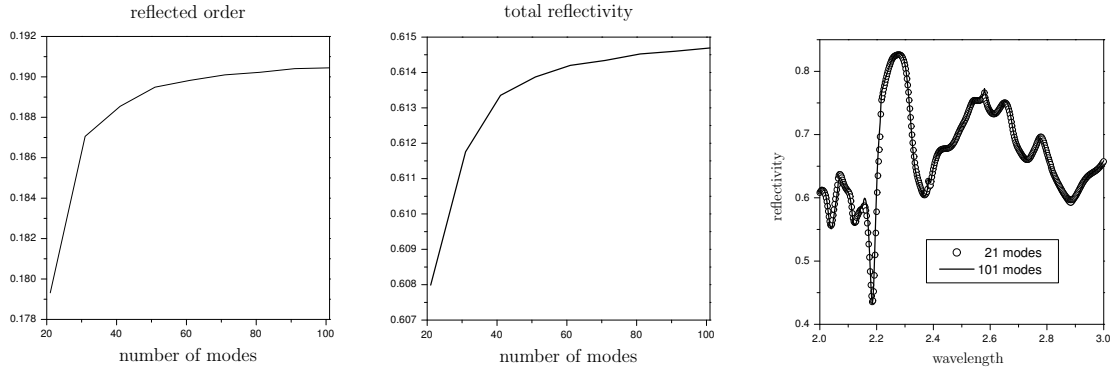


Figure 10.7: The convergence of the main reflected order (left panel) and of the total reflectivity (central panel) when the number of exact eigenfunctions increases. Right panel: total reflectivity as a function of the wavelength (in the arbitrary unit u) for 21 and 101 exact eigenfunctions.

a function of the wavelength $2\pi/(\omega\sqrt{\varepsilon_0\mu_0})$ for 21 and 101 modes. It shows that the exact modal method converges very rapidly since, for $2\pi/(\omega\sqrt{\varepsilon_0\mu_0})$ equal to $2.0u$ and $3.0u$, there are respectively 19 and 13 diffracted orders.

Finally, it is relevant to compare the present results to those obtained through the fictitious sources method. The latter, described in [25, 26, 27, 28, 29, 30], has the ability to solve problems of diffraction by arbitrary-shaped objects and it is moreover well-adapted to perfectly conducting materials. Figure 10.8 shows that the agreement between the two methods is nearly perfect.

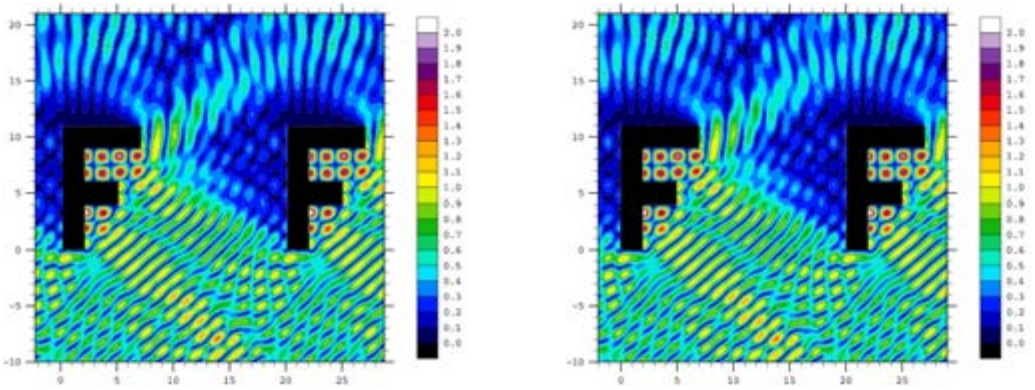


Figure 10.8: Maps of $\log_{10}|E_1|$ *– the electric field along the periodicity direction – using the modal method (left panel) and the fictitious sources method (right panel).*

10.8 Appendix. Calculation of the exact modes and eigenvalues

It is shown here how to determine exactly the eigenvalues and the eigenfunctions of the operator K associated with a lamellar layer in a very general case. An analogous reasoning provides the ones of the operator associated with the others lamellar layers.

From the expression (10.50), every eigenvalue Λ_n of the operator K is either an eigenvalue of K_{ε_1} or K_{μ_1}. So, it is sufficient to determine the set of eigenvalues $\{\Lambda_{\nu_1,n}\,|\,n\in\mathbb{N}\}$ associated with the set of eigenfunctions $\{\phi_{\nu_1,n}\,|\,n\in\mathbb{N}\}$ of the scalar operator K_{ν_1}, with $\nu_1=\varepsilon_1$ or $\nu_1=\mu_1$.

10.8.1 The equation satisfied by the exact eigenvalues

From the expression (10.50), the operator K_{ν_1} is the sum of the two operators $\omega^2\varepsilon_1\mu_1+\partial_1\nu_1^{-1}\partial_1\nu_1$ and ∂_2^2: the first one is an operator of the single variable x_1 and the second is an operator of the single variable x_2. Thus, we can perform a variable separation: every eigenfunction of L_{ν_1} can be written

$$\phi_{\nu_1,n}(x_1,x_2)=\phi_{n_1}^{(1)}(x_1)\,\phi_{n_2}^{(2)}(x_2)\qquad n_1,n_2\in\mathbb{N}\,, \tag{10.78}$$

where $\phi_{n_1}^{(1)}$ and $\phi_{n_2}^{(2)}$ are respectively eigenfunctions of the first and second operators which constitute L_{ν_1}.

In the case of a lamellar grating which is invariant in the direction x_2, this direction of invariance is considered using the Fourier decomposition (10.6). Thus the eigenfunction of ∂_2^2 is just

$$\phi_{n_2}^{(2)}(x_2)=\exp[ik_2x_2]\qquad k_2\in\mathbb{R}\,, \tag{10.79}$$

and the integer n_2 plays no role (integer n will be simply n_1). In the case of woodpile crystals, it is easy to verify that the plane-wave

$$\phi_{n_2}^{(2)}(x_2)=\exp\{i[k_2+2\pi p(n_2)/d_2]x_2/d_2\}\qquad p(n_2)\in\mathbb{Z} \tag{10.80}$$

is an eigenfunction of the operator ∂_2^2 and satisfies the partial Bloch boundary condition (10.8) adapted for the variable x_2. Let $\lambda_{n_2}^{(2)}$ be the associated eigenvalue. Then, from (10.79,10.80),

$$\Lambda_{n_2}^{(2)}=-[k_2+2\pi q(n_2)/d_2]^2\,. \tag{10.81}$$

Note that, for lamellar gratings and eigenfunctions (10.79), the integer $q(n_2)$ is set to zero.

The x_1-dependency of the eigenfunction (10.78) is determined using the usual transfer matrix [31, 32, 33]. Let $\lambda_{n_1}^{(1)}$ be the eigenvalue associated with $\phi_{n_1}^{(1)}$:

$$\left[\omega^2\varepsilon_1\mu_1 + \partial_1\nu_1^{-1}\partial_1\nu_1\right]\phi_{n_1}^{(1)} = \Lambda_{n_1}^{(1)}\phi_{n_1}^{(1)}\,. \tag{10.82}$$

In order to obtain a set of first order differential equations, the following column vector is introduced

$$F_{n_1} = \begin{bmatrix} \nu_1\phi_{n_1}^{(1)} \\ \nu_1^{-1}\partial_1\nu_1\phi_{n_1}^{(1)} \end{bmatrix}. \tag{10.83}$$

Note that, from equation (10.82), the two components of this vector are continuous functions. Now, suppose that the unit cell of the considered lamellar layer is made of J rods of width $w_{1,j}$, permittivity $\varepsilon_{1,j}$ and permeability $\mu_{1,j}$, $j = 1,2,\cdots,J$ (figure 10.9): we denote by $\nu_{1,j}$ the value

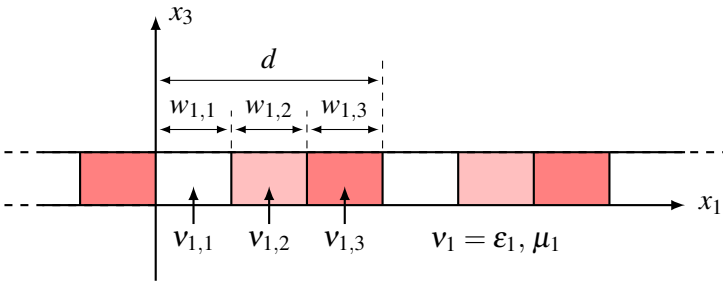


Figure 10.9: A layer made of three rods per unit cell ($J = 3$): the three rods have width $w_{1,j}$, permittivity $\varepsilon_{1,j}$ and permeability $\mu_{1,j}$, $j = 1,2,3$.

of the function ν_1 in the rod j, $j = 1,2,\cdots,$J. Then, from equation (10.82), the vector (10.83) satisfies [33]

$$F_{n_1}(d) = T_1\big(\Lambda_{n_1}^{(1)}\big)F_{n_1}(0)\,, \tag{10.84}$$

where

$$T_1(\Lambda) = T_{1,J}(\Lambda)\;T_{1,J-1}(\Lambda)\;\cdots\;T_{1,1}(\Lambda)\,, \tag{10.85}$$

$$T_{1,j}(\Lambda) = P_{1,j}(\Lambda, w_{1,j})\,, \tag{10.86}$$

$$P_{1,j}(\Lambda, w) = \begin{bmatrix} \cos(\beta_{1,j}w) & \nu_{1,j}\beta_{1,j}^{-1}\sin(\beta_{1,j}w) \\ -\nu_{1,j}^{-1}\beta_{1,j}\sin(\beta_{1,j}w) & \cos(\beta_{1,j}w) \end{bmatrix}, \tag{10.87}$$

$$\beta_{1,j} = \sqrt{\omega^2\varepsilon_{1,j}\mu_{1,j} - \Lambda} \qquad j = 1,2,\cdots,J\,. \tag{10.88}$$

Note that the four elements of each matrix $T_{1,j}$ only depend on $\beta_{1,j}^2$: the expression (10.87) is independent of the definition of the square root (10.88). In addition to (10.84), the vector (10.83) has to satisfy the partial Bloch boundary condition (10.8) for the variable x_1:

$$F_{n_1}(d) = \exp[ik_1 d]F_{n_1}(0)\,. \tag{10.89}$$

The combination of (10.84) and (10.89) implies that $\exp[ik_1 d]$ is an eigenvalue of the matrix $T_1\left(\Lambda_{n_1}^{(1)}\right)$: the equation

$$\det\left\{T_1\left(\Lambda_{n_1}^{(1)}\right) - \exp[ik_1 d]\right\} = 0 \tag{10.90}$$

determines the eigenvalues $\Lambda_{n_1}^{(1)}$. This last equation can be simplified using the fact that $\det T_1 = 1$ (since, from (10.86), $\det T_{1,j} = 1,\ j = 1,2,\cdots,J$): if $\exp[ik_1 d]$ is an eigenvalue of T_1, then $\exp[-ik_1 d]$ is also. Thus, the equation (10.90) is equivalent to

$$\mathrm{tr}T_1\left(\Lambda_{n_1}^{(1)}\right) - 2\cos[k_1 d] = 0\,, \tag{10.91}$$

where $\mathrm{tr}T_1$ is the trace of matrix T_1. Once the eigenvalues $\Lambda_{n_1}^{(1)}$ are determined from (10.91), the associated eigenvectors $\phi_{n_1}^{(1)}$ are also obtained using the transfer matrix [32]: firstly, the eigenvector $F_{n_1}(0)$ in $\mathbb{C}^2$ (associated with the eigenvalue $\exp[ik_1 d]$) of the matrix $T_1\left(\Lambda_{n_1}^{(1)}\right)$ is determined; secondly, the expression of $\phi_{n_1}^{(1)}$ in the rod j can be deduced from

$$F_{n_1}(x_1) = P_{1,j}\left(\Lambda_{n_1}^{(1)}, x_1 - x_{1,j}\right) F_{n_1}(x_{1,j-1})\,, \tag{10.92}$$

where

$$x_{1,0} = 0\,, \quad x_{1,j} = \sum_{q=1}^{j} w_{1,q} \qquad j = 1,2,\cdots,\mathrm{J}\,. \tag{10.93}$$

Finally the eigenvalues of the operator K_{ν_1} are

$$\lambda_{\nu_1,n} = \lambda_{n_1}^{(1)} + \lambda_{n_2}^{(2)}\,, \tag{10.94}$$

whose the two parts are respectively given by (10.91) and (10.81), and the expression of associated eigenvectors is (10.78) whose the two parts are respectively given by (10.92) and (10.80). Concerning the functions of the operators $U = \sigma_1 V$ and $W = \sigma_1^{-1}\partial_1\sigma_1 V$ used in section 10.4 [see equation (10.56)], they are equal to the functions

$$\left(\nu_1\phi_{n_1}^{(1)}\right)(x_1)\,\phi_{n_2}^{(2)}(x_2)\,, \qquad \left(\nu_1^{-1}\partial_1\nu_1\phi_{n_1}^{(1)}\right)(x_1)\,\phi_{n_2}^{(2)}(x_2)\,, \tag{10.95}$$

where n_1 and n_2 are in $\mathbb{N}$, the expression of $\nu_1\phi_{n_1}^{(1)}$ and $\nu_1^{-1}\partial_1\nu_1\phi_{n_1}^{(1)}$ in the rod j can be deduced from (10.83,10.92) and the expression of $\phi_{n_2}^{(2)}$ is given by (10.80).

10.8.2 Real eigenvalues

Here, we suppose that the permittivity and permeability are real positive functions:

$$\varepsilon_1(x_1) \in \mathbb{R}\,, \qquad \varepsilon_+ > \varepsilon_1(x_1) > 0\,; \qquad \mu_1(x_1) \in \mathbb{R}\,, \qquad \mu_+ > \mu_1(x_1) > 0\,. \tag{10.96}$$

Under these conditions, the operator K_{ν_1} is selfadjoint and its eigenvalues are real when the following inner product is used:

$$(\phi, \psi) \longrightarrow \frac{1}{d}\int_0^d \overline{\phi(x_1)}\,\psi(x_1)\,\nu_1(x_1)\,dx_1\,. \tag{10.97}$$

The only difficulty in the numerical determination of the eigenvalues (10.94) is to find the real numbers $\lambda_{n_1}^{(1)}$ which satisfy the transcendental equation (10.91).

Since the numbers $\Lambda_{n_1}^{(1)}$ are eigenvalues of the operator $\omega^2\varepsilon_1\mu_1+\partial_1\nu_1^{-1}\partial_1\nu_1 \leq \omega^2\varepsilon_+\mu_+$, these numbers are on the semi-axis $(-\infty,\omega^2\varepsilon_+\mu_+]$. This property makes their numerical determination easier and provides a way to number them:

$$\omega^2\varepsilon_+\mu_+ \geq \Lambda_{\nu_1,1} \geq \Lambda_{\nu_1,2}\cdots \geq \Lambda_{\nu_1,n} \geq \cdots \tag{10.98}$$

However, two difficulties can occur in this numerical determination. We give herein the solutions we have adopted.

The first difficulty comes from the possibility for two consecutive numbers $\Lambda_{n_1}^{(1)}$ to be very close to each other. Our solution is to use an algorithm which determines the zeros of the function $\mathrm{tr}T_1(\Lambda)-2\cos[k_1 d]$ on the left side of equation (10.91) by taking into account this function together with its derivative with respect to Λ. If two numbers $\Lambda_{n_1}^{(1)}$ are very close one to each other, then the derivative is close to zero. Thus such algorithm needs to determine the function

$$\frac{d}{d\Lambda}\big\{\mathrm{tr}T_1(\Lambda)-2\cos[k_1 d]\big\} = \mathrm{tr}\frac{dT_1}{d\Lambda}(\lambda)\,. \tag{10.99}$$

The expression of the derivative of the matrix T_1 can be deduced from (10.85,10.86):

$$\frac{dT_1}{d\Lambda} = \frac{dT_{1,J}}{d\Lambda}\,T_{1,J-1}\,\cdots\,T_{1,1} + T_{1,J}\,\frac{dT_{1,J-1}}{d\Lambda}\,\cdots\,T_{1,1}+\cdots+T_{1,J}\,T_{1,J-1}\,\cdots\,\frac{dT_{1,1}}{d\Lambda}\,, \tag{10.100}$$

where, for $j=1,2,\cdots,J$, the derivative of matrices

$$\frac{dT_{1,j}}{d\Lambda} = \frac{1}{2}\begin{bmatrix} a_{1,j} & b_{1,j} \\ c_{1,j} & d_{1,j} \end{bmatrix} \tag{10.101}$$

is given by

$$\begin{aligned} a_{1,j} &= w_{1,j}\beta_{1,j}^{-1}\sin[\beta_{1,j}w_{1,j}]\,,\\ b_{1,j} &= \nu_{1,j}\beta_{1,j}^{-3}\sin[\beta_{1,j}w_{1,j}] - \nu_{1,j}w_{1,j}\beta_{1,j}^{-2}\cos[\beta_{1,j}w_{1,j}]\,,\\ c_{1,j} &= \nu_{1,j}^{-1}\beta_{1,j}^{-1}\sin[\beta_{1,j}w_{1,j}] + \nu_{1,j}^{-1}w_{1,j}\cos[\beta_{1,j}w_{1,j}]\,,\\ d_{1,j} &= w_{1,j}\beta_{1,j}^{-1}\sin[\beta_{1,j}w_{1,j}]\,. \end{aligned} \tag{10.102}$$

The second difficulty comes from the possibility of numerical instabilities in the expressions (10.87,10.101) since the numbers $\beta_{1,j}$ (10.88) can have non-vanishing imaginary part. A solution is to multiply the four coefficients of matrices $T_{1,j}$ and their derivative (10.101) by the number

$$N_j = \exp\big[-|\,\mathrm{Im}(\beta_{1,j})|\,w_{1,j}\big] \qquad j=1,2,\cdots,J\,, \tag{10.103}$$

and the term $2\cos[k_1 d]$ wich appears in (10.91) by the product

$$N = N_J N_{J-1}\cdots N_1\,. \tag{10.104}$$

10.8.3 Complex eigenvalues

Here, the permittivity and permeability can take any complex value: $\nu_{1,j}$ is in $\mathbb{C}$, where $\nu_1 = \varepsilon_1, \mu_1$ and $j = 1, 2, \cdots, J$. The operator K_{ν_1} is not selfadjoint and then, its eigenvalues are, in general, in the complex plane. The determination of these complex eigenvalues $\lambda_{n_1}^{(1)}$ which satisfy the equation (10.91) has been intensively studied using different methods [7, 8, 34].

We present here a method similar to the one presented in [8]: the complex eigenvalues are deduced from the real eigenvalues by an analytic continuation. However, our method differs from the one presented in [8] since we make varying the phase of the numbers $\nu_{1,j}$ instead of their imaginary part. We think that it is better to make varying the phase since, from that we have observed, it leaves invariant the generalization to the complex case

$$\mathrm{Re}(\Lambda_{\nu_1,1}) \geq \mathrm{Re}(\Lambda_{\nu_1,2}) \cdots \geq \mathrm{Re}(\Lambda_{\nu_1,p}) \cdots \tag{10.105}$$

of the numbering used when the eigenvalues are real (10.98).

We define for all t in $[0,1]$ the functions

$$\widetilde{\nu}_{1,j}(t) = |\nu_{1,j}| \exp[it \arg(\nu_{1,j})], \tag{10.106}$$

where $\arg(\nu_{1,j})$ is the phase of the complex number $\nu_{1,j}$, $\nu_1 = \varepsilon_1, \mu_1$ and $j = 1, 2, \cdots, J$. Substituting the numbers $\nu_{1,j}$ (where $\nu_1 = \varepsilon_1, \mu_1$) for $\widetilde{\nu}_{1,j}(t)$ in equations (10.85,10.86), we obtain the matrix $\widetilde{T}_1(\Lambda, t)$. For each value of t, we define the numbers $\widetilde{\Lambda}_{n_1}^{(1)}(t)$ which satisfy

$$\mathrm{tr}\widetilde{T}_1\left[\widetilde{\Lambda}_{n_1}^{(1)}(t), t\right] - 2\cos[k_1 d] = 0. \tag{10.107}$$

Then, the numbers $\widetilde{\Lambda}_{n_1}^{(1)}(1)$ are the desired complex eigenvalues $\Lambda_{p_1}^{(1)}$ and the numbers $\widetilde{\Lambda}_{n_1}^{(1)}(0)$ are real eigenvalues which can be determined using the method presented in the previous section 10.8.2. Assuming that $\widetilde{\Lambda}_{n_1}^{(1)}(t)$ are continuous and differentiable functions of t, the complex numbers $\widetilde{\Lambda}_{n_1}^{(1)}(1)$ can be estimated from the numbers $\widetilde{\Lambda}_{n_1}^{(1)}(0)$ by a numerical integration [8] of

$$\frac{d\widetilde{\Lambda}_{n_1}^{(1)}}{dt}(t) = -\frac{\mathrm{tr}(\partial\widetilde{T}_1/\partial\Lambda)\left[\widetilde{\Lambda}_{n_1}^{(1)}(t), t\right]}{\mathrm{tr}(\partial\widetilde{T}_1/\partial t)\left[\widetilde{\Lambda}_{n_1}^{(1)}(t), t\right]}, \tag{10.108}$$

where $\partial\widetilde{T}_1/\partial\Lambda$ is given by substituting the numbers $\nu_{1,j}$ for $\widetilde{\nu}_{1,j}(t)$ in equations (10.100,10.101) and $\partial\widetilde{T}_1/\partial t$ is determined similarly. Finally the obtained estimates of numbers $\widetilde{\Lambda}_{p_1}^{(1)}(1)$ are used to initiate any of the classical methods for the numerical solution of equations [8]. Then, one obtains the desired complex eigenvalues.

In order to eliminate the numerical instabilities, one has to multiply each matrix $T_{1,j}$ and their derivatives by the numbers N_j (10.103) as in the previous section 10.8.2.

10.8.4 Eigenfunctions

From (10.92), the expression of each eigenfunction $\phi_{n_1}^{(1)}$ is given by the coefficients of the column vectors $F_{n_1}(x_{1,j})$, $j = 0, 1, \cdots, J$. On the numerical side, the only difficulty comes from the fact that numerical instabilities in the expression of the transfer matrices (10.86,10.87).

A solution based on the *R*-matrix algorithm (or *S*-matrix) should consist in using the algorithm presented in [35] to obtain the vector $F_{n_1}(x_{1,0})$ (and the vector $F_{n_1}(x_{1,J}) = \exp[k_1 d] F_{n_1}(x_{1,0})$) and then, the algorithm presented in [36, section V] to obtain the vectors $F_{n_1}(x_{1,j})$, $j = 1, 2, \cdots, J-1$. However, we propose to use another solution which benefits of the fact that we deal with 2×2 matrices.

We define the following complex coefficients:

$$\begin{bmatrix} \mathscr{T}_{11}^j & \mathscr{T}_{12}^j \\ \mathscr{T}_{21}^j & \mathscr{T}_{22}^j \end{bmatrix} = T_{1,J}\left(\Lambda_{n_1}^{(1)}\right) T_{1,J-1}\left(\Lambda_{n_1}^{(1)}\right) \cdots T_{1,j}\left(\Lambda_{n_1}^{(1)}\right), \tag{10.109}$$

$$\begin{bmatrix} \tau_{11}^j & \tau_{12}^j \\ \tau_{21}^j & \tau_{22}^j \end{bmatrix} = T_{1,j}\left(\Lambda_{n_1}^{(1)}\right) T_{1,j-1}\left(\Lambda_{n_1}^{(1)}\right) \cdots T_{1,1}\left(\Lambda_{n_1}^{(1)}\right), \tag{10.110}$$

$$\begin{bmatrix} \mathscr{F}_1^j \\ \mathscr{F}_2^j \end{bmatrix} = F_{n_1}(x_{1,j}) \qquad j = 0, 1, \cdots, \mathrm{J}. \tag{10.111}$$

Since $F_{n_1}(x_{1,0})$ is an eigenvector of the matrix $T_1\left(\Lambda_{n_1}^{(1)}\right)$ associated with the eigenvalue $\exp[k_1 d]$), its coefficients satisfy

$$\mathscr{F}_2^0 = -\frac{\mathscr{T}_{11}^J N - \exp[k_1 d] N}{\mathscr{T}_{12}^J N} \mathscr{F}_1^0, \tag{10.112}$$

where the numbers $\mathscr{T}_{11}^J N$ and $\mathscr{T}_{12}^J N$ are obtained by multiplying each coefficient of matrices $T_{1,j}\left(\Lambda_{n_1}^{(1)}\right)$ by the number N_j. The coefficients $\mathscr{F}_1^J$ and $\mathscr{F}_2^J$ are deduced from (10.89,10.112) and then, one can obtain the other coefficients for $j = 1, 2, \cdots, J-1$:

$$\begin{aligned} \mathscr{F}_1^j &= \frac{\mathscr{T}_{22}^{j+1} \tau_{11}^j N}{\mathscr{T}_{21}^{j+1} \tau_{11}^j N + \tau_{21}^j \mathscr{T}_{22}^{j+1} N} \left(\frac{\mathscr{F}_2^J}{\mathscr{T}_{22}^{j+1}} - \frac{\mathscr{F}_2^0}{\tau_{11}^j} \right), \\ \mathscr{F}_2^j &= \frac{\mathscr{T}_{11}^{j+1} \tau_{22}^j N}{\mathscr{T}_{11}^{j+1} \tau_{12}^j N + \tau_{22}^j \mathscr{T}_{12}^{j+1} N} \left(\frac{\mathscr{F}_1^J}{\mathscr{T}_{11}^{j+1}} - \frac{\mathscr{F}_1^0}{\tau_{22}^j} \right), \end{aligned} \tag{10.113}$$

where, as in (10.112), the multiplication by the number N consists in multiplying each coefficient of matrices $T_{1,j}\left(\Lambda_{n_1}^{(1)}\right)$ by the number N_j.

Finally these functions have to be normalized. From the definition (10.97) of the inner product, one has to compute

$$\left\| \phi_{n_1}^{(1)} \right\|_{\nu_1}^2 = \frac{1}{d} \int_0^d \left| \phi_{n_1}^{(1)}(x_1) \right|^2 \nu_1(x_1)\, dx_1 \tag{10.114}$$

when the functions ε_1 and μ_1 have the property (10.96). In the general case (where ε and μ are complex valued functions), one has to use the formalism presented in [9, section 2.3]. It is possible to compute analytically the expression (10.114):

$$\begin{aligned} \left\| \phi_{p_1}^{(1)} \right\|_{\nu_1}^2 = \frac{1}{2 d_{1,1}} \sum_{j=1}^{\mathrm{J}} \frac{w_{1,j}}{\nu_{1,j}} &\left(\left| \mathscr{F}_1^{j-1} \right|^2 + \beta_{1,j}^{-2} \nu_{1,j}^{-2} \left| \mathscr{F}_2^{j-1} \right|^2 \right) \\ &- \beta_{1,j}^{-2} \mathrm{Re}\left(i \overline{\mathscr{F}_1^{j-1}} \mathscr{F}_2^{j-1} - i \overline{\mathscr{F}_1^{j}} \mathscr{F}_2^{j} \right). \end{aligned} \tag{10.115}$$

This expression allows to eliminate the numerical instabilities which can occur from the exponential functions. Note that all the coefficients of matrices defined in sections 10.4 and 10.5 (matrices U, V and W) can be also computed analytically in order to eliminate the numerical instabilities.

10.8.5 The case with infinitely conducting metal

For the sake of simplicity, a single layer made of two rods per unit cell, similar to the one represented on figure 10.4, is considered: it is located between the two horizontal planes defined by equations $x_3 = 0$ and $x_3 = h$, the first rod is made of dielectric material with dielectric constant $\varepsilon_{1,1} = \varepsilon_a$ and width $w_{1,1} = a$, and the second rod is made of infinitely conducting metal (its width is $w_{1,2} = d - a$). Thus, defining the characteristic function

$$\Psi_a(x_1) = \begin{cases} 1\,, & 0 \le x_1 + p\,d \le a \\ 0\,, & a < x_1 + p\,d < d \end{cases}\,, \qquad p \in \mathrm{Z}\,, \tag{10.116}$$

the set of equations (10.72) restricted to the domain $0 \le x_3 \le h$ becomes:

$$\boldsymbol{E} = \Psi_a(\omega\varepsilon_a)^{-1}\boldsymbol{\nabla}\times\boldsymbol{H}\,, \quad \boldsymbol{H} = (\omega\mu_0)^{-1}\boldsymbol{\nabla}\times\boldsymbol{E}\,. \tag{10.117}$$

After the Floquet and Fourier decompositions (see section 10.2), the equations (10.117) above becomes

$$\widehat{\boldsymbol{E}} = \Psi_a(\omega\varepsilon_a)^{-1}\boldsymbol{\nabla}_{k_2}\times\widehat{\boldsymbol{H}}\,, \quad \widehat{\boldsymbol{H}} = (\omega\mu_0)^{-1}\boldsymbol{\nabla}_{k_2}\times\widehat{\boldsymbol{E}}\,, \tag{10.118}$$

where $\nabla_{k_2}\times$ is the curl operator with the partial derivation ∂_2 replaced by ik_2. For all fixed Bloch wave vector k_1, let $\mathscr{H}(k_1)$ be the Hilbert space of functions which satisfy the two conditions (10.7) and (10.8), *i.e.* the space square integrable function on the domain $[0,d]$ with the partial Bloch boundary condition. The combination of the square integrability (10.7) together with the equations (10.118) imposes that: the tangential components of $\widehat{\boldsymbol{E}}$ are continuous at all the interfaces separating dielectrics and infinitely conducting metal [since $\widehat{\boldsymbol{H}} = (\omega\mu_0)^{-1}\boldsymbol{\nabla}_{k_2}\times\widehat{\boldsymbol{E}}$ everywhere]; and the tangential components of $\widehat{\boldsymbol{H}}$ are continuous at all the interfaces separating dielectrics. More precisely, in the present case, the metallic rod imposes the conditions:

$$\begin{aligned} &\widehat{E}_1(x_1,k_1,k_2,x_3) = \widehat{E}_2(x_1,k_1,k_2,x_3) = 0\,, && a \le x_1 \le d\,, \quad x_3 = 0,h\,; \\ &\widehat{E}_2(x_1,k_1,k_2,x_3) = \widehat{E}_3(x_1,k_1,k_2,x_3) = 0\,, && 0 \le x_3 \le h, \quad x_1 = 0,a\,. \end{aligned} \tag{10.119}$$

From (10.118) and (10.119), the components $\widehat{E}_2$ and $\widehat{E}_3$ of the electric field are continuous functions of the variable x_1 and satisfy the equation

$$\left[\partial_3^2 + L_a\right]\widehat{E}_j = 0\,, \quad L_a = \omega^2\Psi_a\varepsilon_a\mu_0 - k_2^2 + \partial_1^2\,, \quad j = 2,3\,. \tag{10.120}$$

where L_a is acting on the Hilbert space $\mathscr{H}_a(k_1) \subset \mathscr{H}(k_1)$ defined by $\mathscr{H}_a(k_1) = \{\,\phi = \Psi_a\psi \,|\, \psi \in \mathscr{H}(k_1)\,,\, \phi(0) = \phi(a) = 0\,\}$. Since the Fourier decomposition (10.6) has been performed in the present case, the solely x_1-dependence is considered (the x_2-dependence is the same than in section 10.8.1). Let $\{\,\phi_{a,n_1}\,|\,n_1 \in \mathrm{N}\,\}$ be the set of the eigenfunctions of L_a and $\{\,\lambda_{a,n_1}\,|\,n_1 \in \mathrm{N}\,\}$ the associated eigenvalues:

$$L_a\phi_{a,n_1} = \lambda_{a,n_1}\,\phi_{a,n_1}\,, \quad n_1 \in \mathrm{N}\,. \tag{10.121}$$

The expression of these eigenfunctions and the associated eigenvalues is

$$\phi_{a,n_1} : x_1 \longmapsto \sqrt{\frac{2}{a}}\Psi_a(x_1)\sin[n_1\pi x_1/a]\,, \quad \lambda_{a,n_1} = \omega^2\varepsilon_a\mu_0 - k_2^2 - \left(\frac{n_1\pi}{a}\right)^2. \tag{10.122}$$

Note that, for $n_1 = 0$, the function $\phi_{a,0}$ is not an eigenfunction of the operator L_a since it is the null function. We include it because it is more convenient for the next calculations. Developing the components $\widehat{E}_2$ and $\widehat{E}_3$ on this orthonormal set of eigenfunctions, we obtain from (10.120) the following (formal) expression:

$$\widehat{E}_j(x_3) = \sum_{n_1\in\mathrm{N}} \phi_{a,n_1}\left[\widehat{E}_j^{(a,n_1)}(0)\cos\left(\sqrt{\lambda_{a,n_1}}\,x_3\right) + \left(\partial_3\widehat{E}_j^{(a,n_1)}\right)(0)\frac{\sin\left(\sqrt{\lambda_{a,n_1}}\,x_3\right)}{\sqrt{\lambda_{a,n_1}}}\right], \quad j = 2,3\,, \tag{10.123}$$

where the coefficients $\widehat{E}_j^{(a,n_1)}(0)$ and $\left(\partial_3\widehat{E}_j^{(a,n_1)}\right)(0)$ are respectively the projection uppon the functions ϕ_{a,n_1} of $\widehat{E}_j$ and $\partial_3\widehat{E}_j$,

$$\begin{aligned}
\widehat{E}_j^{(a,n_1)}(x_3) &= \int_{[0,d]} dx_1\phi_{a,n_1}(x_1)\widehat{E}_j(x_1,x_3)\,,\\
\left(\partial_3\widehat{E}_j^{(a,n_1)}\right)(x_3) &= \int_{[0,d]} dx_1\phi_{a,n_1}(x_1)\left(\partial_3\widehat{E}_j\right)(x_1,x_3)\,, \quad j = 2,3\,,
\end{aligned} \tag{10.124}$$

taken at $x_3 = 0$.

From (10.118), the electric field satisfies $\boldsymbol{\nabla}\cdot\widehat{\boldsymbol{E}} = 0$. Then, the expression of the first component $\widehat{E}_1$ of the electric field can be deduced from the expression (10.123) of the other two components: $\partial_1\widehat{E}_1 = -ik_2\widehat{E}_2 - \partial_3\widehat{E}_3$. In particular, its x_1-dependence can be developed on the eigenfunctions of the operator

$$L'_a = \omega^2\varepsilon_a\mu_0 - k_2^2 + \partial_1^2\,, \tag{10.125}$$

acting on the Hilbert space $\mathscr{H}'_a(k_1) \subset \mathscr{H}(k_1)$ defined by $\mathscr{H}'_a(k_1) = \{\,\phi = \Psi_a\psi \,|\, \psi \in \mathscr{H}(k_1)\,, (\partial_1\phi)(0) = (\partial_1\phi)(a) = 0\,\}$. Let $\{\,\phi'_{a,n_1}\,|\,n_1 \in \mathrm{N}\,\}$ be the set of the eigenfunctions of L'_a and $\{\,\lambda_{a,n_1}\,|\,n_1 \in \mathrm{N}\,\}$ the associated eigenvalues:

$$L'_a\phi'_{a,n_1} = \lambda_{a,n_1}\,\phi'_{a,n_1}\,, \quad n_1 \in \mathrm{N}\,. \tag{10.126}$$

The expression of these eigenfunctions is

$$\begin{aligned}
&\phi'_{a,0} : x_1 \longmapsto \sqrt{\frac{1}{a}}\Psi_a(x_1)\,,\\
&\phi'_{a,n_1} : x_1 \longmapsto \sqrt{\frac{2}{a}}\Psi_a(x_1)\cos[n_1\pi x_1/a]\,, \quad n_1 \in \mathrm{N}\setminus\{0\}\,.
\end{aligned} \tag{10.127}$$

Note that the numbering of the eigenfunctions of L_a and L'_a is done such that, for all n_1 in N, they are associated with the same eigenvalues given by (10.122).

Finally, the modal basis associated with the magnetic field is deduced from the equation $\widehat{\boldsymbol{H}} = (\omega\mu_0)^{-1}\nabla_{k_2}\times\widehat{\boldsymbol{E}}$ (10.118). The x_1-dependence of the component $\widehat{H}_1$ can be developed on the eigenfunctions of the operator L_a (10.122) while the x_1-dependence of the components $\widehat{H}_2$ and $\widehat{H}_3$ can be developed on the eigenfunctions of the operator L'_a (10.127).

References:

[1] M. G. Moharam and T. K. Gaylord, "Rigorous coupled-waves analysis of metallic surface-relief grating," J. Opt. Soc. Am. A **3**, 1780–1787 (1986).

[2] L. Li, "New formulation of the Fourier modal method for crossed surface-relief gratings," J. Opt. Soc. Am. A **14**, 2758–2767 (1997).

[3] L. Li, "Justification of matrix truncation in the modal methods of diffraction gratings," J. Opt. A: Pure Appl. Opt. **1**, 531–536 (1999).

[4] B. Gralak, M. de Dood, G. Tayeb, S. Enoch, and D. Maystre, "Theoretical study of photonic band gaps in woodpile crystals," Phys. Rev. E **67**, 066 601 (2003).

[5] L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, "The dielectric lamellar diffraction grating," Optica acta **28**, 413–428 (1981).

[6] L. C. Botten, M. S. Craig, R. C. McPhedran, J. L. Adams, and J. R. Andrewartha, "The finitely conducting lamellar diffraction grating," Optica acta **28**, 1087–1102 (1981).

[7] L. C. Botten, M. S. Craig, and R. C. McPhedran, "Highly conducting lamellar diffraction grating," Optica acta **28**, 1103–1106 (1981).

[8] G. Tayeb and R. Petit, "On the numerical study of deep conducting lamellar diffraction grating," Optica Acta **31**, 1361–1365 (1984).

[9] L. Li, "A modal analysis of lamellar diffraction gratings in conical mountings," Journal of Modern Optics **40**, 553–573 (1993).

[10] E. Silberstein, P. Lalanne, J.-P. Hugonin, and Q. Cao, "Use of grating theories in integrated optics," J. Opt. Soc. Am. B **18**, 2865 (2001).

[11] A. Tip, A. Moroz, and J. M. Combes, "Band structure of absorptive photonic crystals," J. Phys. A: Math. Gen. **33**, 6223–6252 (2000).

[12] B. Gralak and S. Guenneau, "Transfer matrix method for point sources radiating in classes of negative refractive index materials with 2n-fold antisymmetry," Waves in Random and Complex Media **17**, 581 (2007).

[13] L. Li, "Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings," J. Opt. Soc. Am. A **13**, 1024–1035 (1996).

[14] L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," J. Opt. Soc. Am. A **13**, 1870–1876 (1996).

[15] T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff, "Extraordinary optical transmission through subwavelength hole arrays," Nature **391**, 667–669 (1998).

[16] S. Enoch, M. Nevière, E. Popov, and R. Reinisch, "Enhanced light transmission by hole arrays," J. Opt. A: Pure Appl. Opt. **4**, S83–S87 (2002).

[17] S. Enoch, G. Tayeb, P. Sabouroux, N. Guérin, and P. Vincent, "A Metamaterial for Directive Emission," Phys. Rev. Lett. **89**, 213 902 (2002).

[18] P. Andrew and W. L. Barnes, "," Phys. Rev. B **64**, 125 405 (2001).

[19] J. Kalkman, C. Strohhöfer, B. Gralak, and A. Polman, "Surface plasmon polariton modified emission of Erbium in a metallodielectric grating," Appl. Phys. Lett. **83**, 30 (2003).

[20] Y. De Wilde, F. Formanek, R. Carminati, B. Gralak, P.-A. Lemoine, K. Joulain, J.-P. Mulet, Y. Chen, and J.-J. Greffet, "Thermal radiation scanning tunnelling microscopy," Nature **444**, 740–743 (2006).

[21] M. C. K. Wiltshire, J. B. Pendry, I. R. Young, D. J. Larkman, D. J. Gilderdale, and J. V. Hajnal, "Microstructured Magnetic Materials for RF Flux Guides in Magnetic Resonance Imaging," Science **291**, 849–851 (2001).

[22] R. A. Shelby, D. R. Smith, and S. Schultz, "Experimental verification of a negative index of refraction," Science **292**, 77–79 (2001).

[23] B. Gralak, R. Pierre, G. Tayeb, and S. Enoch, "Solutions of Maxwell's equations in presence of lamellar gratings including infinitely conducting metal," J. Opt. Soc. Am. A **25**, 3099 (2008).

[24] Z.-Y. Li and K.-M. Ho, "Analytic modal solution to light propagation through layer-by-layer metallic photonic crystals," Phys. Rev. B **67**, 165 104 (2003).

[25] C. Hafner, The Generalized Multipole Technique for Computational Electromagnetics (Artech House Books, Boston, 1990).

[26] G. Tayeb, "The method of fictitious sources applied to diffraction gratings," Special issue on Generalized Multipole Techniques (GMT) of Applied Computational Electromagnetics Society Journal **9**, 90–100 (1994).

[27] D. Maystre, M. Saillard, and G. Tayeb, Scattering (Academic Press, London, 2001).

[28] D. Kaklamani and H. Anastassiu, "Aspects of the method of auxiliary sources (MAS) in computational electromagnetics," IEEE Ant. and Prop. Magazine **44** (2002).

[29] G. Tayeb and S. Enoch, "Combined Fictitious Sources - Scattering Matrix method," J. Opt. Soc. Am. A **21**, 1417–1423 (2004).

[30] G. Benelli, S. Enoch, and G. Tayeb, "Modelling of a single object embedded in a layered medium," Journal of Modern Optics **54**, 871–879 (2007).

[31] M. Reed and B. Simon, Methods of Modern Mathematical Physics, Vol. IV: Analysis of Operators (Academic Press, 1978).

[32] A. Figotin and V. Gorentsveig, "Localized electromagnetic waves in a layered periodic dielectric medium with a defect," Phys. Rev. B **58**, 180–188 (1998).

[33] D. Felbacq, B. Guizal, and F. Zolla, "Wave propagation in one-dimensional photonic crystals," Optics Communications **152**, 119–126 (1998).

[34] S.-E. Sandström, G. Tayeb, and R. Petit, "Lossy multistep lamellar gratings in conical diffraction mountings: an exact eigenfunction solution," Journal of Electromagnetic Waves and Applications **7**, 631–649 (1993).

[35] B. Gralak, S. Enoch, and G. Tayeb, "From scattering or impedance matrices to Bloch modes of photonic crystals," J. Opt. Soc. Am. A **19**, 1547–1554 (2002).

[36] D. M. Whittaker and I. S. Culshaw, "Scattering-matrix treatment of patterned multilayer photonic structures," Phys. Rev. B **60**, 2610–2618 (1999).

Chapter 11:
Homogenization Techniques for Periodic Structures

Sebastien Guenneau,

Richard Craster,

Tryfon Antonakakis,

Elizabeth Skelton,

Kirill Cherednichenko, and

Shane Cooper

Table of Contents:

Chapter 11

Homogenization Techniques for Periodic Structures

Sebastien Guenneau[1], Richard Craster[2], Tryfon Antonakakis[2,3], Elizabeth Skelton[2], Kirill Cherednichenko[4] and Shane Cooper[4]

[1] *CNRS, Aix-Marseille Université, École Centrale Marseille, Institut Fresnel, 13397 Marseille Cedex 20, France, sebastien.guenneau@fresnel.fr*

[2] *Department of Mathematics, Imperial College London, United Kingdom, r.craster@imperial.ac.uk,*

[3] *CERN, Geneva, Switzerland,tryfon.antonakakis09@imperial.ac.uk,*

[4] *Cardiff School of Mathematics, Cardiff University, United Kingdom, cherednichenko@cardiff.ac.uk, coopersa@cf.ac.uk.*

11.1 Introduction

In this chapter we describe a selection of mathematical techniques and results that suggest interesting links between the theory of gratings and the theory of homogenization, including a brief introduction to the latter. By no means do we purport to imply that homogenization theory is an exclusive method for studying gratings, neither do we aim to be exhaustive in our choice of topics within the subject of homogenization. Our preferences here are motivated most of all by our own latest research, and by our outlook to the future interactions between these two subjects. We have also attempted, in what follows, to contrast the "classical" homogenization (Section 11.1.2), which is well suited for the description of composites as we have known them since their advent until about a decade ago, and the "non-standard" approaches, high-frequency homogenization (Section 11.2) and high-contrast homogenization (Section 11.3), which have been developing in close relation to the study of photonic crystals and metamaterials, which exhibit properties unseen in conventional composite media, such as negative refraction allowing for super-lensing through a flat heterogeneous lens, and cloaking, which considerably reduces the scattering by finite size objects (invisibility) in certain frequency range. These novel electromagnetic paradigms have renewed the interest of physicists and applied mathematicians alike in the theory of gratings [1].

11.1.1 Historical survey on homogenization theory

The development of theoretical physics and continuum mechanics in the second half of the 19th and first half of the 20th century has motivated the question of justifying the macrosopic view of physical phenomena (at the scales visible to the human eye) by "upscaling" the implied microscopic rules for particle interaction at the atomic level through the phenomena at the intermediate, "mesoscopic", level (from tenths to hundreds of microns). This ambition has led to an extensive worldwide programme of research, which is still far from being complete as of now. Trying to give a very crude, but more or less universally applicable, approximation of the aim of this extensive activity, one could say that it has to do with developing approaches to averaging out in some way material properties at one level with the aim of getting a less detailed, but almost equally precise, description of the material response. Almost every word in the last sentence needs to be clarified already, and this is essentially the point where one

could start giving an overview of the activities that took place during the years to follow the great physics advances of a century ago. Here we focus on the research that has been generally referred to as the theory of homogenization, starting from the early 1970s. Of course, even at that point it was not, strictly speaking, the beginning of the subject, but we will use this period as a kind of reference point in this survey.

The question that a mathematician may pose in relation to the perceived concept of "averaging out" the detailed features of a heterogeneous structure in order to get a more homogeneous description of its behaviour is the following: suppose that we have the simplest possible linear elliptic partial differential equation (PDE) with periodic coefficients of period $\eta > 0$. What is the asymptotic behaviour of the solutions to this PDE as $\eta \to 0$? Can a boundary-value problem be written that is satisfied by the leading term in the asymptotics, no matter what the data unrelated to material properties are? Several research groups became engaged in addressing this question about four decades ago, most notably those led by N. S. Bakhvalov, E. De Giorgi, J.-L. Lions, V. A. Marchenko, see [2], [3], [4], [5] for some of the key contributions of that period. The work of these groups has immediately led to a number of different perspectives on the apparently basic question asked above, which in part was due to the different contexts that these research groups had had exposure to prior to dealing with the issue of averaging. Among these are the method of multiscale asymptotic expansions (also discussed later in this chapter), the ideas of compensated compactness (where the contribution by L. Tartar and F. Murat [6], [7] has to be mentioned specifically), the variational method (also known as the "Γ-convergence"). These approaches were subsequently applied to various contexts, both across a range of mathematical setups (minimisation problems, hyperbolic equations, problems with singular boundaries) and across a number of physical contexts (elasticity, electromagnetism, heat conduction). Some new approaches to homogenization appeared later on, too, such as the method of two-scale convergence by G. Nguetseng [8] and the periodic unfolding technique by D. Cioranescu, A. Damlamian and G. Griso [9]. Established textbooks that summarise these developments in different time periods, include, in addition to the already cited book [4], the monographs [10], [11], [12], and more recently [13]. The area that is perhaps worth a separate mention is that of stochastic homogenization, where some pioneering contributions were made by S. M. Kozlov [14], G. C. Papanicolaou and S. R. S. Varadhan [15], and which has in recent years been approached with renewed interest.

A specific area of interest within the subject of homogenization that has been rapidly developing during the last decade or so is the study of the behaviour of "non-classical" periodic structures, which we understand here as those for which compactness of bounded-energy solution sequences fails to hold as $\eta \to 0$. The related mathematical research has been strongly linked to, and indeed influenced by, the parallel development of the area of metamaterials and their application in physics, in particular for electromagnetic phenomena. Metamaterials can be roughly defined as those whose properties at the macroscale are affected by higher-order behaviour as $\eta \to 0$. For example, in classical homogenization for elliptic second-order PDE one requires the leading ("homogenised solution") and the first-order ("corrector") terms in the η-power-series expansion of the solution in order to determine the macroscopic properties, which results in a limit of the same type as the original problem, where the solution flux ("stress" in elasticity, "induction" in electromagnetics, "current" in electric conductivity, "heat flux" in heat conduction) depends on the solution gradient only ("strain" in elasticity, "field" in electromagnetics, "voltage" in electric conductivity, "temperature gradient" in heat condiction). If, however, one decides for some reason, or is forced by the specific problem setup, to include higher-order terms as well, they are likely to have to deal with an asymptotic limit of a different

type for small η, which may, say, include second gradients of the solution in its constitutive law. One possible reason for the need to include such unusual effects is the non-uniform (in η) ellipticity of the original problems or, using the language of materials science, the high-contrast in the material properties of the given periodic structure. Perhaps the earliest mathematical example of such degeneration is the so-called "double-porosity model", which was first considered by G. Allaire [16] and T. Arbogast, J. Douglas, U. Hornung [17] in the early 1990s. A detailed analysis of the properties of double-porosity models, including their striking spectral behaviour did not appear until the work [18] by V. V. Zhikov. We discuss the double-porosity model and its properties in more detail in Section 11.3.

Before moving on to the next section, it is important to mention one line of research within the homogenization area that has had a significant rôle in terms of application of mathematical analysis to materials, namely the subject of periodic singular structures (or "multi-structures", see [19]). While this subject is clearly linked to the general analysis of differential operators on singular domains (see [20]), there has been a series of works that develop specifically homogenization techniques for periodic structures of this kind (also referred to as "thin structures" in this context), *e.g.* [21], [22]. It turns out that overall properties of such materials are similar to those of materials with high contrast. In the same vein, it is not difficult to see that compactness of bounded-energy sequences for problems on periodic thin structures does not hold (unless the sequence in question is suitably rescaled), which leads to the need for non-classical, higher-order, techniques in their analysis.

11.1.2 Multiple scale method: Homogenization of microstructured fibers

Let us consider a doubly periodic grating of pitch η and finite extent such as shown in Fig.11.1. An interesting problem to look at is that of transverse electric (TE) modes— when the magnetic field has the form $(0,0,H)$— propagating within a micro-structured fiber with infinite conducting walls. Such an eigenvalue problem is known to have a discrete spectrum: we look for eigenfrequencies ω and associated eigenfields H such that:

$$(\mathscr{P}_\eta): \begin{cases} -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(\varepsilon_{ij}^{-1}(\frac{\mathbf{x}}{\eta})\frac{\partial H(\mathbf{x})}{\partial x_j}\right) = \omega^2\mu_0\varepsilon_0 H(\mathbf{x}) & \text{in } \Omega_f\ , \\ \varepsilon_{ij}^{-1}(\frac{\mathbf{x}}{\eta})\frac{\partial H(\mathbf{x})}{\partial x_i} n_j = 0 & \text{on } \partial\Omega_f\ , \end{cases}$$

where we use the convention $\mathbf{x} = (x_1, x_2)$, $\partial\Omega_f$ denotes the boundary Ω_f, and $\mathbf{n} = (n_1, n_2)$ is the normal to the boundary. Here, $\varepsilon_0\mu_0 = c^{-2}$ where c is the speed of light in vacuum and we assume that matrix coefficients of relative permittivity $\varepsilon_{ij}(\mathbf{y})$, with $i, j = 1, 2$, are real, symmetric (with the convention $\mathbf{y} = (y_1, y_2)$), of period 1 (in y_1 et y_2) and satisfy:

$$M|\,\boldsymbol{\xi}\,|^2 \geq \varepsilon_{ij}(\mathbf{y})\xi_i\xi_j \geq m|\,\boldsymbol{\xi}\,|^2\ ,\ \forall \boldsymbol{\xi} \in \mathbb{R}^2\ ,\ \forall \mathbf{y} \in Y = [0,1]^2\ , \tag{11.1}$$

where $|\,\boldsymbol{\xi}\,|^2 = (\xi_1^2 + \xi_2^2)$, for given strictly positive constants M and m. This condition is met for all conventional dielectric media[1].

[1]When the periodic medium is assumed to be isotropic, $\varepsilon_{ij}(y) = \varepsilon(\mathbf{y})\delta_{ij}$, with the Kronecker symbol $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. For instance, (11.1) has typically the bounds $M = 13$ and $m = 1$ in optics. One class of problems where this condition (11.1) is violated (the bound below, to be more precise) is considered in Section 11.3 on high-contrast homogenization.

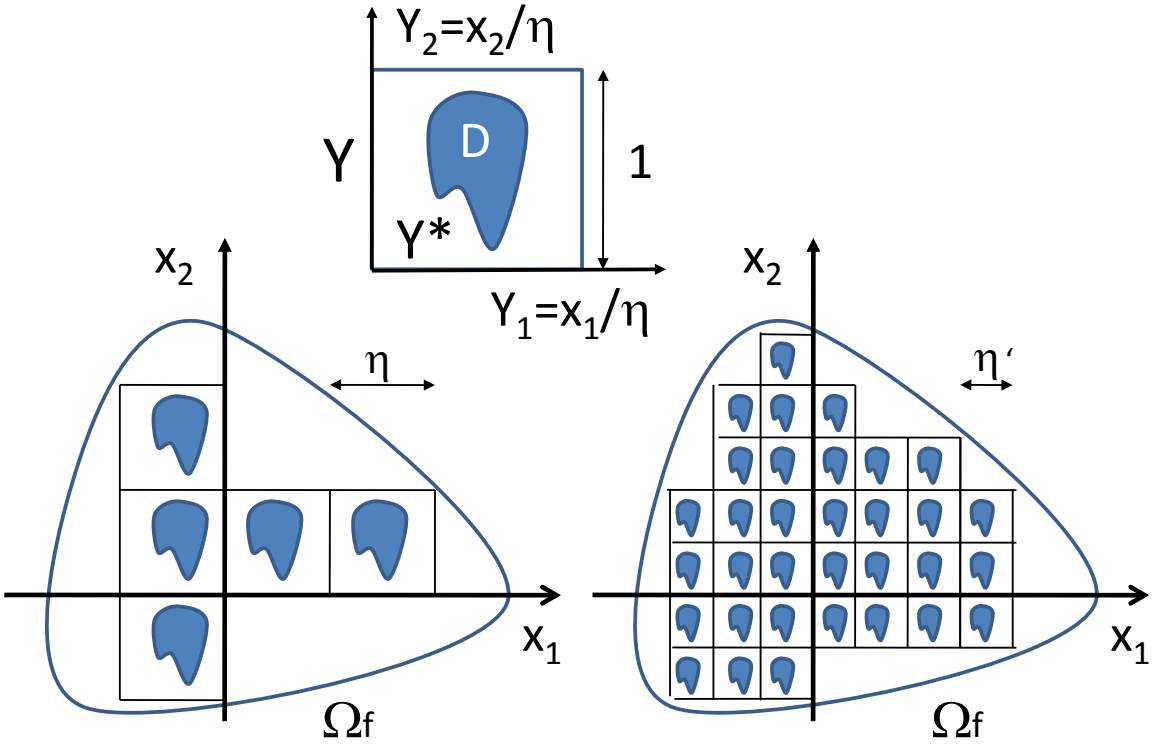


Figure 11.1: A diagram of the homogenization process: when the parameter η gets smaller ($\eta < \eta'$), the number of cells inside the fixed domain Ω_f becomes larger. When $\eta \ll 1$, Ω_f is filled with a large number of small cells, and can thus be considered as an effective (or homogenized) medium. Such a medium is usually described by anisotropic parameters depending upon the resolution of auxiliary ("unit cell") problems set on the rescaled microcopic cell Y which typically contains one inclusion D.

We can recast $(\mathscr{P}_\eta)$ as follows:

$$-\frac{\partial}{\partial x_i}\sigma^i(H(\mathbf{x})) = \frac{\omega^2}{c^2}H(\mathbf{x})$$

with

$$\sigma^i(H(\mathbf{x})) = \varepsilon_{ij}^{-1}\left(\frac{\mathbf{x}}{\eta}\right)\frac{\partial H(\mathbf{x})}{\partial x_j}\,.$$

The multiscale method relies upon the following ansatz:

$$H = H_0(\mathbf{x}) + \eta H_1(\mathbf{x},\mathbf{y}) + \eta^2 H_2(\mathbf{x},\mathbf{y}) + \ldots \tag{11.2}$$

where $H_i(\mathbf{x},\mathbf{y})$, $i = 1,2,\ldots$ is a periodic function of period Y in $\mathbf{y}$.
In order to proceed with the asymptotic algorithm, one needs to rescale the differential operator as follows

$$\frac{\partial H}{\partial x_i} = \left(\frac{\partial H_0}{\partial z_i} + \frac{\partial H_1}{\partial y_i}\right) + \eta\left(\frac{\partial H_1}{\partial z_i} + \frac{\partial H_2}{\partial y_i}\right) + \ldots \tag{11.3}$$

where $\partial/\partial z_i$ stands for the partial derivative with respect to the ith component of the macroscopic variable $\mathbf{x}$.
It is useful to set

$$\sigma^i(H) = \sigma_0^i + \eta\sigma_1^i + \eta^2\sigma_2^i + ...$$

what makes (11.3) more compact.
Collecting coefficients sitting in front of the same powers of η, we obtain:

$$\sigma_0^i(H) = \varepsilon_{ij}^{-1}(\mathbf{y})\left(\frac{\partial H_0}{\partial z_i} + \frac{\partial H_1}{\partial y_i}\right)$$

$$\sigma_1^i(H) = \varepsilon_{ij}^{-1}(\mathbf{y})\left(\frac{\partial H_1}{\partial z_i} + \frac{\partial H_2}{\partial y_i}\right)$$

and so forth, all terms being periodic in $\mathbf{y}$ of period 1.
Upon inspection of problem $(\mathscr{P}_\eta)$, we gather that

$$-\left(\frac{1}{\eta}\frac{\partial}{\partial y_i} + \frac{\partial}{\partial z_i}\right)\left(\sigma_0^i + \eta\sigma_1^i + ...\right) = \frac{\omega^2}{c^2}H(\mathbf{x}) + ...$$

so that at order η^{-1}

$$(\mathscr{A}) : -\frac{\partial}{\partial y_i}\sigma_0^i = 0\ ,$$

and at order η^0

$$(\mathscr{H}) : -\frac{\partial}{\partial z_i}\sigma_0^i - \frac{\partial}{\partial y_i}\sigma_1^i = \frac{\omega^2}{c^2}H_0\ .$$

(the equations corresponding to higher orders in η will not be used here).
Let us show that $(\mathscr{H})$ provides us with an equation (known as the homogenized equation) associated with the macroscopic behaviour of the microstructured fiber. Its coefficients will be obtained thanks to $(\mathscr{A})$ which is an auxiliary problem related to the microscopic scale. We will therefore be able to compute H_0 and H_1 thus, in particular, the first terms of H and σ^i.
In order to do so, let us introduce the mean on Y, which we denote $< . >$, which is an operator acting on the function g of the variable $\mathbf{y}$:

$$< g >= \frac{1}{\mid Y \mid}\int\int_Y g(y_1, y_2)dy_1dy_2\ ,$$

where $\mid Y \mid$ is the area of the cell Y.
Applying the mean to both sides of $(\mathscr{H})$, we obtain:

$$< (\mathscr{H}) >: -\frac{\partial}{\partial z_i} < \sigma_0^i > - < \frac{\partial}{\partial y_i}\sigma_1^i >= \frac{\omega^2}{c^2}H_0 < 1 >\ ,$$

where we have used the fact that $< . >$ commutes with $\partial/\partial z_i$.
Moreover, invoking the divergence theorem, we observe that

$$< \frac{\partial}{\partial y_i}\sigma_1^i >= \frac{1}{\mid Y \mid}\int\int_Y \frac{\partial}{\partial y_i}\sigma_1^i(\mathbf{y})d\mathbf{y} = \frac{1}{\mid Y \mid}\int_{\partial Y}\sigma_1^i(\mathbf{y})n_i ds\ ,$$

where $\mathbf{n} = (n_1, n_2)$ is the unit outside normal to ∂Y of Y. This normal takes opposite values on opposite sides of Y, hence the integral over ∂Y vanishes.

Altogether, we obtain:

$$< (\mathscr{H}) >: -\frac{\partial}{\partial z_i} < \sigma_0^i >= \frac{\omega^2}{c^2} H_0 \ ,$$

which only involves the macroscopic variable x and partial derivatives $\partial/\partial z_i$ with respect to the macroscopic variable. We now want to find a relation between $< \sigma_0 >$ and the gradient in $\mathbf{x}$ of H_0. Indeed, we have seen that

$$\sigma_0^i(H) = \varepsilon_{ij}^{-1}(\mathbf{y}) \left(\frac{\partial H_0}{\partial z_j} + \frac{\partial H_1}{\partial y_j} \right) \ ,$$

which from $(\mathscr{A})$ leads to

$$(\mathscr{A}1) : -\frac{\partial}{\partial y_i} \left(\varepsilon_{ij}^{-1}(\mathbf{y}) \frac{\partial H_1}{\partial y_j} \right) = \left(\frac{\partial H_0}{\partial z_j} \right) \left(\frac{\partial}{\partial y_i} \varepsilon_{ij}^{-1}(\mathbf{y}) \right) \ .$$

We can look at $(\mathscr{A}1)$ as an equation for the unknown $H_1(\mathbf{x},\mathbf{y})$, periodic of period Y in $\mathbf{y}$ and parametrized by $\mathbf{x}$. Such an equation is solved up to an additive constant. In addition to that, the parameter $\mathbf{x}$ is only involved via the factor $\partial H_0/\partial z_j$. Hence, by linearity, we can write the solution $H_1(\mathbf{x},\mathbf{y})$ as follows:

$$H_1(\mathbf{x},\mathbf{y}) = \frac{\partial H_0(\mathbf{x})}{\partial z_j} w^j(\mathbf{y}) \ ,$$

where the two functions $w^j(\mathbf{y})$, $j = 1,2$ are solutions to $(\mathscr{A}1)$ corresponding to $\partial H_0/\partial z_j(\mathbf{x})$, $j = 1,2$ equal to unity with the other ones being zero, that is solutions to:

$$(\mathscr{A}2) : -\frac{\partial}{\partial y_i} \left(\varepsilon_{ij}^{-1}(\mathbf{y}) \frac{\partial w^k}{\partial y_j} \right) = \delta_{jk} \left(\frac{\partial}{\partial y_i} \varepsilon_{ij}^{-1}(\mathbf{y}) \right) \ ,$$

with $w^k(\mathbf{y})$, $k = 1,2$ periodic functions in $\mathbf{y}$ of period Y [2].
Since the functions $w^k(\mathbf{y})$ are known, we note that

$$\sigma_i^0(\mathbf{x},\mathbf{y}) = \varepsilon_{ij}^{-1}(\mathbf{y}) \left(\frac{\partial H_0}{\partial z_j} + \frac{\partial H_1}{\partial y_j} \right) = \varepsilon_{ij}^{-1}(\mathbf{y}) \left(\frac{\partial H_0}{\partial z_j} + \frac{\partial H_0}{\partial z_k} \frac{\partial w^k(\mathbf{y})}{\partial y_j} \right) \ ,$$

which can be written as

$$\sigma_0^i(\mathbf{x},\mathbf{y}) = \left(\varepsilon_{ik}^{-1}(\mathbf{y}) + \varepsilon_{ij}^{-1}(\mathbf{y}) \frac{\partial w^k(\mathbf{y})}{\partial y_j} \right) \frac{\partial H_0(\mathbf{x})}{\partial z_k} \ .$$

Lets us now apply the mean to both sides of this equation. We obtain:

$$< \sigma_0^i > (\mathbf{x}) = \varepsilon_{\text{hom},ik}^{-1} \frac{\partial H_0(\mathbf{x})}{\partial z_k} \ ,$$

which can be recast as the following homogenized problem:

[2] We note that $(\mathscr{A}2)$ are two equations which merely depend upon $\varepsilon_{ij}^{-1}(\mathbf{y})$, that is on the microscopic properties of the periodic medium. The two functions w^k (defined up to an additive constant) can be computed once for all, independently of Ω_f.

Figure 11.2: Potentials V_x (left) and V_y (right): The unit cell contains an elliptic inclusion of relative permittivity ($\varepsilon = 4.0 + 3i$) with minor and major axis $a = 0.3$ and $b = 0.4$ in silica ($\varepsilon = 1.25$).

$$(\mathscr{P}_0): \begin{cases} -\sum_{i,k=1}^{2} \frac{\partial}{\partial x_i}\left(\varepsilon^{-1}_{\text{hom},ik}\frac{\partial H_0(\mathbf{x})}{\partial x_k}\right) = \omega^2\mu_0\varepsilon_0 H_0(\mathbf{x}) & ,\text{in } \Omega_f\ , \\ \varepsilon^{-1}_{\text{hom},ik}(\frac{\mathbf{x}}{\eta})\frac{\partial H_0(\mathbf{x})}{\partial x_i}n_k = 0 & ,\text{on } \partial\Omega_f\ , \end{cases}$$

where $\varepsilon^{-1}_{\text{hom},ik}$ denote the coefficients of the homogenized matrix of permittivity given by:

$$\varepsilon^{-1}_{\text{hom},ik} = \frac{1}{|\,Y\,|}\int\!\!\int_Y \left(\varepsilon^{-1}_{ik}(\mathbf{y}) + \varepsilon^{-1}_{ij}(\mathbf{y})\frac{\partial w^k(\mathbf{y})}{\partial y_j}\right) d\mathbf{y}\ . \tag{11.4}$$

As an illustrative example for this homogenized problem, we consider a microstructured waveguide consisting of a medium with relative permittivity $\varepsilon = 1.25$ with elliptic inclusions (of minor and major axes 0.3 cm and 0.4 cm respectively) with center to center spacing $d = 0.1cm$ with an infinite conducting boundary *i.e.* Neumann boundary conditions in the TE polarization.

We use the COMSOL MULTIPHYSICS finite element package to solve the annex problem and we find that $[\varepsilon_{\text{hom}}]$ from (11.4) writes as [26]

$$\begin{pmatrix} 1.9296204 & -1.0533083\,10^{-16} \\ -44.417444\,10^{-18} & 2.1127643 \end{pmatrix},$$

with $<\varepsilon>_Y = 2.2867255$. The off diagonal terms can be neglected.

If we assume that the transverse propagating modes in the metallic waveguide have a small propagation constant $\gamma \ll 1$, the above mathematical model describes accurately the physics. We show in Fig. 11.3 a comparison between two TE modes of the microstructured waveguide and its associated anisotropic homogenized counterpart. Both eigenfrequencies and eigenfields match well (note that we use the waveguide terminology wavenumber $k = \sqrt{\omega^2/c^2 - \gamma^2}$).

11.1.3 The case of one-dimensional gratings: Application to invisibility cloaks

There is a case of particular importance for applications in grating theory: that of a periodic multilayered structure. Let us assume that the permittivity of this medium is $\varepsilon = \alpha$ in white layers and β in yellow layers, as shown in Fig. 11.4.
Equation $(\mathscr{A}2)$ takes the form:

$$(\mathscr{A}3): -\frac{d}{dy}\left(\varepsilon^{-1}(y)\frac{dw}{\partial y}\right) = \left(\frac{d}{dy}\varepsilon^{-1}(y)\right)\ ,$$

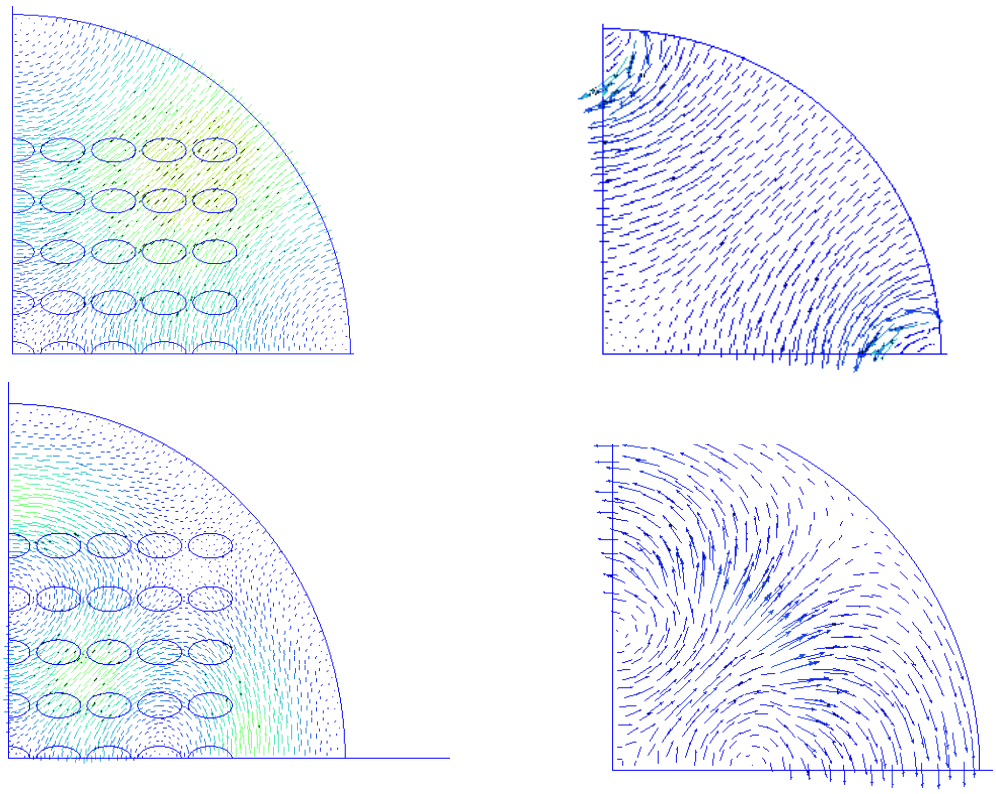


Figure 11.3: Comparison between transverse electric fields TE_{21} and TE_{31} of a microstructured metallic waveguide for a propagation constant $\gamma = 0.1cm^{-1}$ (wavenumbers $k = 0.7707cm^{-1}$ and $k = 0.5478cm^{-1}$ respectively), see left panel, with the TE_{21} and TE_{31} modes of the corresponding homogenized anisotropic metallic waveguide for $\gamma = 0.1cm^{-1}$ ($k = 0.7607cm^{-1}$ and $k = 0.5201cm^{-1}$, where $k = \sqrt{\omega^2/c^2 - \gamma^2} = \sqrt{\omega^2\varepsilon_0\mu_0 - \gamma^2}$ were obtained from the computation of eigenvalues ω of homogenized problem $(\mathscr{P}_0)$), see right panel.

with $w(y)$, periodic function in y of period 1.
We deduce that

$$-\frac{dw}{dy} = 1 + C\varepsilon(y) \, .$$

Noting that $\displaystyle\int_Y \frac{dw}{dy} = w(1) - w(0) = 0$, this leads to

$$\int_Y (1 + C\varepsilon(y))\, dy = 0 \, .$$

Since $\mid Y \mid = 1$, we conclude that

$$C = - < \varepsilon >^{-1} \, .$$

The homogenized permittivity takes the form:

$$\begin{aligned} \varepsilon_{\text{hom}}^{-1} &= \frac{1}{\mid Y \mid} \int_Y \left(\varepsilon^{-1}(y) + \varepsilon^{-1}(y) \frac{dw(y)}{dy} \right) dy \\ &= < \varepsilon^{-1}(y) > - < \varepsilon^{-1}(y) + C > \\ &= < \varepsilon^{-1}(y) > - < \varepsilon^{-1}(y) > + < < \varepsilon(y) >^{-1} > = < \varepsilon(y) >^{-1} \, . \end{aligned}$$

We note that if we now consider the full operator i.e. we include partial derivatives in y_1 and y_2, the anisotropic homogenized permittivity takes the form:

$$\varepsilon_{\text{hom}}^{-1} = \begin{pmatrix} < \varepsilon(y)^{-1} > & 0 \\ 0 & < \varepsilon(y) >^{-1} \end{pmatrix} \, ,$$

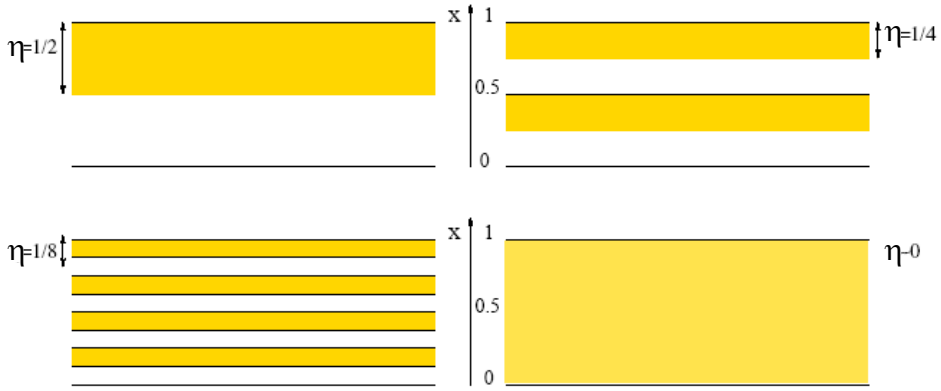


Figure 11.4: Schematic of homogenization process for a one-dimensional grating with homogeneous dielectric layers of permittivity α and β in white and yellow regions. When η tends to zero the number of layers tends to infinity, and their thicknesses vanish, in such a way that the width of the overall stack remains constant.

as the only contribution for $\varepsilon^{-1}_{\text{hom},11}$ is $1/\mid Y \mid \int_Y \varepsilon^{-1}(y)\,dy$.
As an illustrative example of what artificial anisotropy can achieve, we propose the design of an invisibility cloak. For this, let us assume that we have a multilayered grating with periodicity along the radial axis. In the coordinate system (r,θ), the homogenized permittivity clearly has the same form as above. If we want to design an invisibility cloak with an alternation of two homogeneous isotropic layers of thicknesses d_A and d_B and permittivities α, β, we then need to use the formula

$$\frac{1}{\varepsilon_r} = \frac{1}{1+\eta}\left(\frac{1}{\alpha}+\frac{\eta}{\beta}\right), \quad \varepsilon_\theta = \frac{\alpha+\eta\beta}{1+\eta},$$

where $\eta = d_B/d_A$ is the ratio of thicknesses for layers A and B and $d_A + d_B = 1$.

We now note that the coordinate transformation $r' = R_1 + r\frac{R_2-R_1}{R_2}$ can compress a disc $r < R_2$ into a shell $R_1 < r < R_2$, provided that the shell is described by the following anisotropic heterogeneous permittivity [27] $\underline{\underline{\varepsilon}}^{\text{cloak}}$ (written in its diagonal basis):

$$\varepsilon_r^{\text{cloak}} = \left(\frac{R_2}{R_2-R_1}\right)^2\left(\frac{r'-R_1}{r'}\right)^2, \quad \varepsilon_\theta^{\text{cloak}} = \left(\frac{R_2}{R_2-R_1}\right)^2, \tag{11.5}$$

where R_1 and R_2 are the interior and the exterior radii of the cloak. Such a metamaterial can be approximated using the formula (11.5), as first proposed in [28], which leads to the multilayered cloak shown in Fig. 11.5.

11.2 High-frequency homogenization

Many of the features of interest in photonic crystals [44, 45], or other periodic structures, such as all-angle negative refraction [46, 47, 48, 49] or ultrarefraction [50, 51] occur at high frequencies

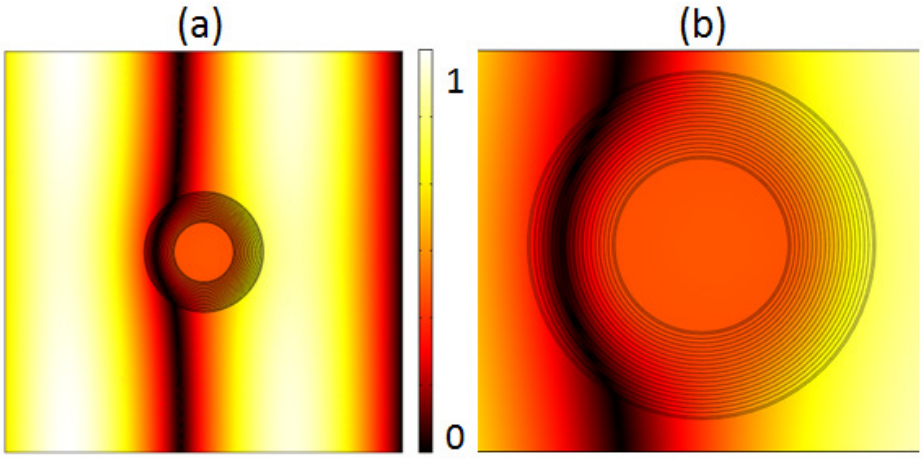


Figure 11.5: Propagation of a plane wave of wavelength $7\ 10^{-7}m$ *(red in the visible spectrum) from the left on a multilayered cloak of inner radius* $R_1 = 1.5\ 10^{-8}m$ *and outer radius* $R_2 = 3\ 10^{-8}m$*, consisting of 20 homogeneous layers of equal thickness and of respective relative permittivities* $1680.70, 0.25, 80.75, 0.25, 29.39, 0.25, 16.37, 0.25, 10.99, 0.25, 8.18, 0.25, 6.50, 0.25, 5.40, 0.25, 4.63, 0.25, 4.06, 0.25$ *in vacuum. Importantly, one layer in two has the same permittivity.*

where the wavelength and microstructure dimension are of similar orders. Therefore the conventional low-frequency classical homogenisation clearly fails to capture the essential physics and a different approach to distill the physics into an effective model is required. Fortunately a high frequency homogenisation (HFH) theory as developed in [37] is capable of capturing features such as AANR and ultra-refraction [52] for some model structures. Somewhat tangentially, there is an existing literature in the analysis community on Bloch homogenisation [53, 54, 55, 56], that is related to what we call high frequency homogenisation. There is also a flourishing literature on developing homogenised elastic media, with frequency dependent effective parameters, based upon periodic media [38]. There is therefore considerable interest in creating effective continuum models of microstructured media that break free from the conventional low frequency homogenisation limitations.

11.2.1 High Frequency Homogenization for Scalar Waves

Waves propagating through photonic crystals and metamaterials have proven to show different effects depending on their frequency. The homogenization of a periodic material is not unique. The effective properties of a periodic medium change depending on the vibration modes within its cells. The dispersion diagram structure can be considered to be the identity of such a material and provides the most important information regarding group velocities, band-gaps of disallowed propagation frequency bands, Dirac cones and many other interesting effects. The goal of a homogenization theory is to provide an effective homogeneous medium that is equivalent, in the long scale, to the initial non-homogeneous medium composed of a short-scale periodic, or other microscale, structure. This was achieved initially using the classical theory of homogenization [4, 34, 11, 35, 36] and yields an intuitively obvious result that the effective medium's properties consist of simple averages of the original medium's properties. This is valid so long as the wavelength is very large compared to the size of the cells (here we focus on periodic media created by repeating cells). For shorter wavelengths of the order of a cell's length a more general theory has been developed [37] that also recovers the results of the classical homogenization theory. For clarity we present high frequency homogeniaztion (HFH) by means of an

illustrative example and consider a two-dimensional lattice geometry for TE or TM polarised electromagnetic waves. With harmonic time dependence, $\exp(-i\Omega t)$ (assumed understood and henceforth suppressed), the governing equation is the scalar Helmholtz equation,

$$\nabla^2 u + \Omega^2 u = 0, \tag{11.6}$$

where u represent E_Z and H_Z, for TM and TE polarised electromagnetic waves respectively, and $\Omega^2 = n^2\omega^2/c^2$. In our example the cells are square and each square cell of length $2l$ contains a circular hole and the filled part of the cell has constant non-dimensionalized properties. The boundary conditions on the hole's surface, namely the boundary ∂S_2, depend on the polarisation and are taken to be either of Dirichlet or Neumann type. This approach assumes infinite conducting boundaries which is a good approximation for micro-waves. We adopt a multiscale approach where l is the small length scale and L is a large length scale and we set $\eta = l/L \ll 1$ to be the ratio of these scales. The two length scales let us introduce the following two independent spatial variables, $\xi_i = x_i/l$ and $X_i = x_i/L$. The cell's reference coordinate system is then $-1 < \xi < 1$. By introducing the new variables in equation (11.6) we obtain,

$$u(\mathbf{X},\boldsymbol{\xi})_{,\xi_i\xi_i} + \Omega^2 u(\mathbf{X},\boldsymbol{\xi}) + 2\eta u(\mathbf{X},\boldsymbol{\xi})_{,\xi_i X_i} + \eta^2 u(\mathbf{X},\boldsymbol{\xi})_{,X_i X_i} = 0. \tag{11.7}$$

We now pose an ansatz for the field and the frequency,

$$\begin{aligned} u(\mathbf{X},\boldsymbol{\xi}) &= u_0(\mathbf{X},\boldsymbol{\xi}) + \eta u_1(\mathbf{X},\boldsymbol{\xi}) + \eta^2 u_2(\mathbf{X},\boldsymbol{\xi}) + \ldots, \\ \Omega^2 &= \Omega_0^2 + \eta\Omega_1^2 + \eta^2\Omega_2^2 + \ldots \end{aligned} \tag{11.8}$$

In this expansion we set Ω_0 to be the frequency of standing waves that occur in the perfectly periodic setting. By substituting equations (11.8) into equation (11.7) and grouping equal powers of ε through to second order, we obtain a hierarchy of three ordered equations:

$$u_{0,\xi_i\xi_i} + \Omega_0^2 u_0 = 0, \tag{11.9}$$

$$u_{1,\xi_i\xi_i} + \Omega_0^2 u_1 = -2u_{0,\xi_i X_i} - \Omega_1^2 u_0, \tag{11.10}$$

$$u_{2,\xi_i\xi_i} + \Omega_0^2 u_2 = -u_{0,X_i X_i} - 2u_{1,\xi_i X_i} - \Omega_1^2 u_1 - \Omega_2^2 u_0. \tag{11.11}$$

These equations are solved as in [40, 37] and hence the description is brief.

The asymptotic expansions are taken about the standing wave frequencies that occur at the corners of the irreducible Brillouin zone depicted in Fig. 11.6. It should be noted that not all structured cells will have the usual symmetries of a square, as in Fig. 11.6(a) where there is no reflexion symmetry from the diagonals. As a consequence the usual triangular region ΓXM does not always represent the irreducible Brillouin zone and the square region ΓMXN should be used instead. Also paths that cross the irreducible Brillouin zone have proven to yield interesting effects namely along the path MX' for large circular holes [39].

The subsequent asymptotic development considers small perturbations about the points Γ, X and M so that the boundary conditions of u on the outer boundaries of the cell, namely ∂S_1, read,

$$u|_{\xi_i=1} = \pm u|_{\xi_i=-1} \quad \text{and} \quad u_{,\xi_i}|_{\xi_i=1} = \pm u_{,\xi_i}|_{\xi_i=-1}, \tag{11.12}$$

where the $+,-$ stand for periodic and anti-periodic conditions respectively: the standing waves occur when these conditions are met. The conditions on ∂S_2 are either of Dirichlet or Neumann

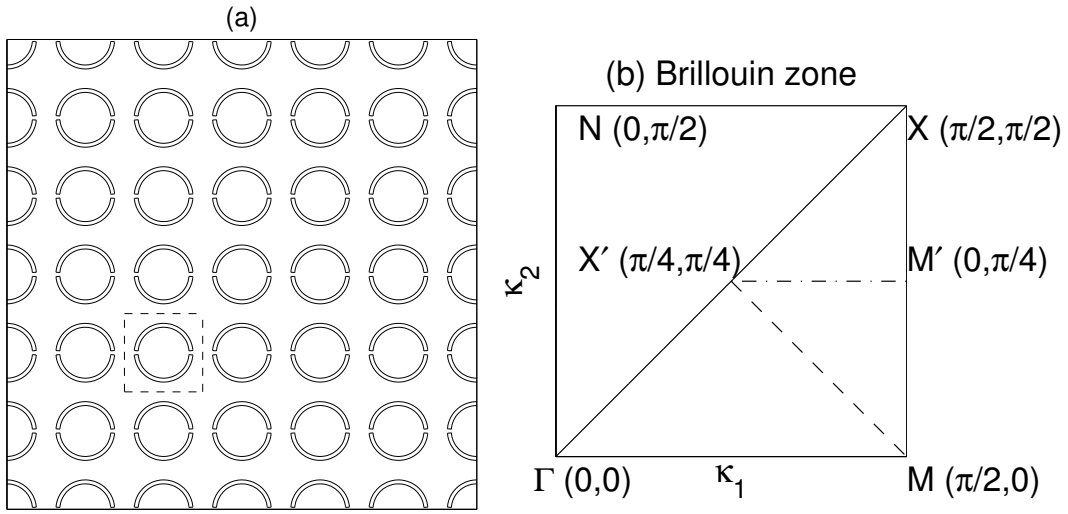


Figure 11.6: Panel (a) An infinite square array of split ring resonators with the elementary cell shown as the dashed line inner square. Panel (b) shows the irreducible Brillouin zone, in wavenumber space, used for square arrays in perfectly periodic media based around the elementary cell shown of length 2l ($l = 1$ in (b)). Figure reproduced from Proceedings of the Royal Society [40].

type. The theory that follows is similar for both boundary condition cases, but the latter one is illustrated herein. Neumann boudary condition on the hole's surface or equivalently electromagnetic waves in TE polarization yield,

$$\frac{\partial u}{\partial \mathbf{n}} = u_{,x_i} n_i|_{\partial S_2} = 0. \tag{11.13}$$

which in terms of the two-scales and $u_i(\mathbf{X}, \boldsymbol{\xi})$ become

$$U_{0,\xi_i} n_i = 0, \quad (U_0 f_{0,X_i} + u_{1,\xi_i}) n_i = 0, \quad (u_{1,X_i} + u_{2,\xi_i}) n_i = 0. \tag{11.14}$$

The solution of the leading order equation is by introducing the following separation of variables $u_0 = f_0(\mathbf{X}) U_0(\boldsymbol{\xi}; \Omega_0)$. It is obvious that $f_0(\mathbf{X})$, which represents the behaviour of the solution in the long scale, is not set by the leading order equation and the resulting eigenvalue problem is solved on the short-scale for Ω_0 and U_0 representing the standing wave frequencies and the associated cell's vibration modes respectively. To solve the first order equation (11.10) we take the integral over the cell of the product of equation (11.10) with U_0 minus the product of equation (11.9) with u_1/f_0 and this yields $\Omega_1 = 0$. It then follows to solve for $u_1(\mathbf{X}, \boldsymbol{\xi}) = f_{0,X_i}(\mathbf{X}) U_{1_i}(\boldsymbol{\xi})$ where the vector $\mathbf{U}_1$ is found as in [40]. By invoking a similar solvability condition for the second order equation we obtain a second order PDE for $f_0(\mathbf{X})$,

$$\begin{aligned} T_{ij} f_{0,X_i X_j} + \Omega_2^2 f_0 = 0 \quad \text{where,} \\ T_{ij} = \frac{t_{ij}}{\int\int_S U_0^2 dS} \quad \text{for} \quad i, j = 1, 2 \end{aligned} \tag{11.15}$$

entirely on the long scale with the coefficients T_{ij} containing all the information of the cell's dynamical response and the tensor t_{ij} represents dynamical averages of the properties of the medium. For Neumann boundary conditions on ∂S_2 its formulation reads,

$$t_{ii} = \int\int_S U_0^2 dS + \int\int_S (U_{1_i,\xi_i} U_0 - U_{1_i} U_{0,\xi_i}) dS \quad \text{for} \quad i = 1 \text{ or } 2, \tag{11.16}$$

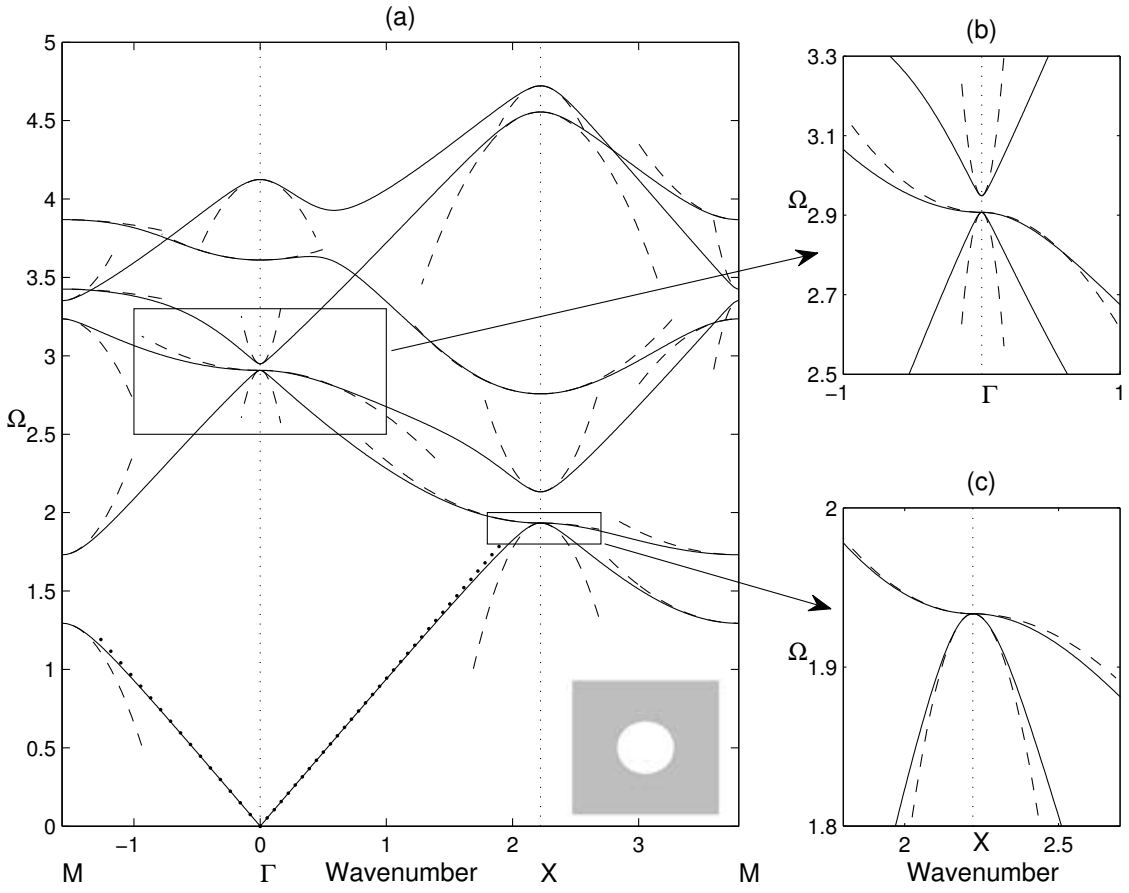


Figure 11.7: The dispersion diagram for a doubly periodic array of square cells with circular inclusions, of radius 0.4*, free at their inner boundaries shown for the irreducible Brillouin zone of Fig. 11.6. The dispersion curves are shown in solid lines and the asymptotic solutions from the high frequency homogenization theory are shown in dashed lines. Figure reproduced from Proceedings of the Royal Society [40].*

$$t_{ij} = \int\int_S (U_{1_j,\xi_i} U_0 - U_{1_j} U_{0,\xi_i}) dS \quad \text{for} \quad i \neq j. \tag{11.17}$$

Note that there is no summation over repeated indexes for t_{ii}. The tensor depends on the boundary conditions of the holes and has a different form if Dirichlet type conditions are applied on ∂S_2.

The PDE for f_0 has several uses, and can be verified by re-creating asymptotically the dispersion curves for a perfect lattice system. One important result of equation (11.15) is its use in the expansion of Ω namely in equation (11.8). In order to obtain Ω_2 as a function of the Bloch wavenumbers we use the Bloch boundary conditions on the cell to solve for $f_0(\mathbf{X}) = \exp(i\kappa_j X_j/\eta)$, where $\kappa_j = K_j - d_j$ with $d_j = 0, \pi/2, -\pi/2$ depending on the location in the Brillouin zone. The asymptotic dispersion relation now reads,

$$\Omega \sim \Omega_0 + \frac{T_{ij}}{2\Omega_0} \kappa_i \kappa_j. \tag{11.18}$$

Equation (11.18) yields the behaviour of the dispersion curves asymptotically around the standing wave frequencies that are naturally located at the edge points of the Brillouin zone. Fig. 11.8 illustrates the asymptotic dispersion curves for the first six dispersion bands of a square cell geometry with circular holes.

An assumption in the development of equation (11.18) is that the standing wave frequencies are isolated. But one can clearly see in Fig. 11.7 that this is not the case for third standing wave frequency at point Γ as well as for the second standing wave frequency at point X. A small alteration to the theory [40] enables the computation of the dispersion curves at such points by setting,

$$u_0 = f_0^{(l)}(\mathbf{X})U_0^{(l)}(\boldsymbol{\xi};\Omega_0) \tag{11.19}$$

where we sum over the repeated superscripts (l). Proceeding as before, we multiply equation (11.10) by $U_0^{(m)}$, substract $u_1((U_{0,\xi_i}^{(m)})_{\xi_i} + \Omega_0^2 U_0^{(m)})$ then integrate over the cell to obtain,

$$\left(\frac{\partial}{\partial X_j}\mathbf{A}_{jml} + \Omega_1^2\mathbf{B}_{ml}\right)\hat{f}_0^{(l)} = 0, \quad \text{for} \quad m = 1,2,\ldots,p \tag{11.20}$$

Ω_1 is not necessarily zero, and

$$\mathbf{A}_{jml} = \int\int_S (U_0^{(m)}U_{0,\xi_j}^{(l)} - U_{0,\xi_j}^{(m)}U_0^{(l)})dS, \quad \mathbf{B}_{ml} = \int\int U_0^{(l)}U_0^{(m)}dS. \tag{11.21}$$

There is now a system of coupled partial differential equations for the $f_0^{(l)}$ and, provided $\Omega_1 \neq 0$, the leading order behaviour of the dispersion curves near the Ω_0 is now linear (these then form Dirac cones).

For the perfect lattice, we set $f_0^{(l)} = \hat{f}_0^{(l)}\exp(i\kappa_j X_j/\eta)$ and obtain the following index equations,

$$(i\frac{\kappa_j}{\eta}\mathbf{A}_{jml} + \Omega_1^2\mathbf{B}_{ml})\hat{f}_0^{(l)} = 0, \quad \text{for} \quad \mathrm{m} = 1,2,...,\mathrm{p} \tag{11.22}$$

The system of equation (11.22) can be written simply as,

$$\mathbf{C}\hat{\mathbf{F}}_0 = 0, \tag{11.23}$$

with $\mathbf{C}_{ll} = \Omega_1^2\mathbf{B}_{ll}$ and $\mathbf{C}_{ml} = i\kappa_j\mathbf{A}_{jml}/\eta$ for $l \neq m$. One must then solve for $\Omega_1^2 = \pm\sqrt{\alpha_{ij}\kappa_i\kappa_j}/\eta$ when the determinant of $\mathbf{C}$ vanishes and insert the result in,

$$\Omega \sim \Omega_0 \pm \frac{1}{2\Omega_0}\sqrt{\alpha_{ij}\kappa_i\kappa_j}. \tag{11.24}$$

If the Ω_1 are zero one must go to the next order.

11.2.1.1 Repeated eigenvalues: quadratic asymptotics

If Ω_1 is zero, $u_1 = f_{0,X_k}^{(l)}U_{1_k}^{(l)}$ (we again sum over all repeated (l) superscripts) and we advance to second order using (11.11). Taking the difference between the product of equation (11.11) with $U_0^{(m)}$ and $u_2(U_{0,\xi_i\xi_i} + \Omega_0^2 U_0)$ and then integrating over the elementary cell gives

$$\begin{aligned} &f_{0,X_iX_i}^{(l)}\textstyle\int\int_S U_0^{(m)}U_0^{(l)}dS + f_{0,X_kX_j}^{(l)}\int\int_S(U_0^{(m)}U_{1_k,\xi_j}^{(l)} - U_{0,\xi_j}^{(m)}U_{1_k}^{(l)})dS \\ &\quad + \Omega_2^2 f_0^{(l)}\textstyle\int\int_S U_0^{(m)}U_0^{(l)}dS = 0, \quad \text{for} \quad m = 1,2,...,p \end{aligned} \tag{11.25}$$

as a system of coupled PDEs. The above equation is presented more neatly as

$$f_{0,X_iX_i}^{(l)}\mathbf{A}_{ml} + f_{0,X_kX_j}^{(l)}\mathbf{D}_{kjml} + \Omega_2^2 f_0\mathbf{B}_{ml} = 0, \quad \text{for} \quad m = 1,2,...,p. \tag{11.26}$$

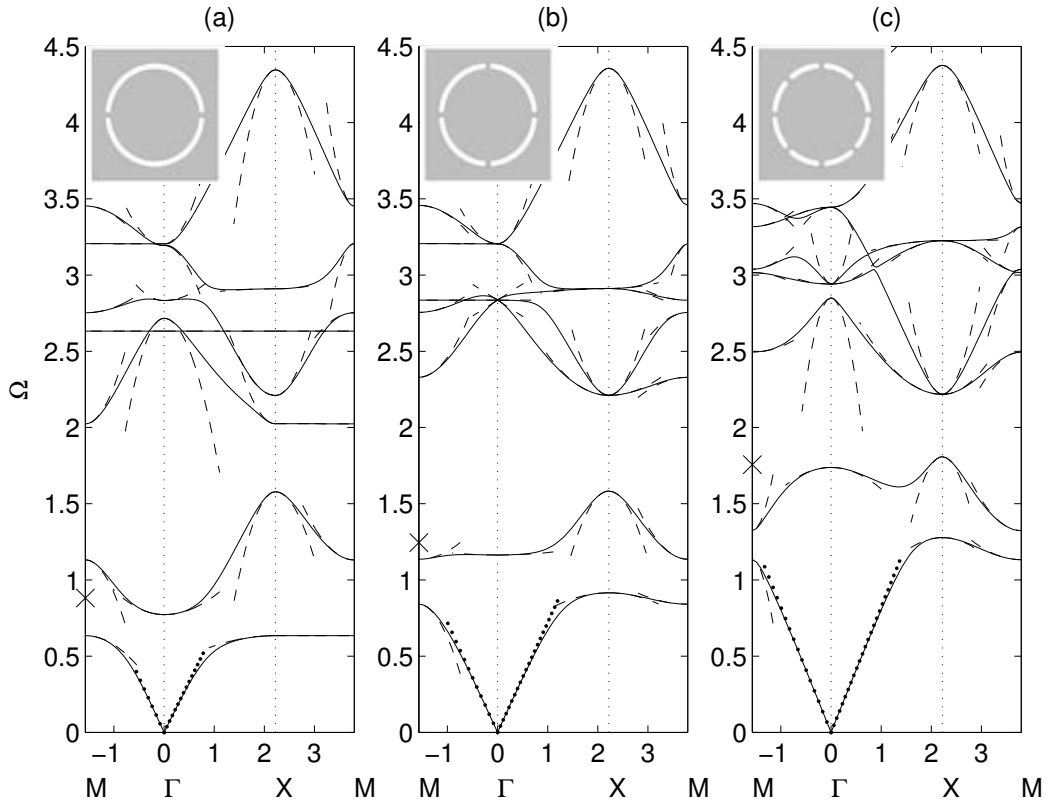


Figure 11.8: The dispersion diagrams for a doubly periodic array of square cells with split ring inclusions, free at their inner boundaries shown for the irreducible Brillouin zone of Fig. 11.6. The dispersion curves are shown in solid lines and the asymptotic solutions from the high frequency homogenization theory are shown in dashed lines. Figure reproduced from Proceedings of the Royal Society [40].

For the Bloch wave setting, using $f_0^{(l)}(\mathbf{X}) = \hat{f}_0^{(l)} \exp(i\kappa_j X_j/\eta)$ we obtain the following system,

$$\left(-\frac{\kappa_i\kappa_i}{\eta^2}\mathbf{A}_{ml} - \frac{\kappa_k\kappa_j}{\eta^2}\mathbf{D}_{kjml} + \Omega_2^2\mathbf{B}_{ml}\right)\hat{f}_0^{(l)} = 0, \quad \text{for} \quad m = 1,2,...,p \tag{11.27}$$

and this determines the asymptotic dispersion curves.

11.2.1.2 The classical long wave zero frequency limit

The current theory simplifies if one enters the classical long wave, low frequency limit where $\Omega^2 \sim O(\varepsilon^2)$ as U_0 becomes uniform, and without loss of generality is set to be unity, over the elementary cell. The final equation is again (11.15) where the tensor t_{ij} simplifies to

$$t_{ii} = \int\int_S dS + \int\int_S U_{1_i,\xi_i} dS, \quad t_{ij} = \int\int_S U_{1_j,\xi_i} dS \quad \text{for} \quad i \neq j \tag{11.28}$$

(with no summation over repeated suffices in this equation) and $T_{ij} = t_{ij}/\int\int_S dS$.

11.2.2 Illustrations for Tranverse Electric Polarized Waves

Let us now turn to some illustrative examples. We present in Fig. 11.8 the TE polarization waves for three types of SRR's (Split Ring Resonator's). Equation (11.15) represents the wave

propagation in the effective medium. It is noticable that the T_{ij} coefficients depend on the standing wave frequency and that T_{11} is not necessarily equal to T_{22} in order to yield an anisotropic effective medium for each separate frequency. Near some of the standing wave frequencies the anisotropy effects are very pronounced and well explained by the no longer elliptic equation (11.15).

In the above equations U_{1_i} is a solution of,

$$U_{1_j,\xi_i\xi_i} = 0, \tag{11.29}$$

with boundary conditions $(f_{0,X_i} + u_{1,\xi_i})n_i = 0$ on the hole boundary. If the medium is homogeneous as it is in the illustrative examples herein, equation (11.29) is the same as that for U_0, but with different boundary conditions. The specific boundary conditions for U_{1_j} are

$$U_{1_j,\xi_i} n_i = -n_j \quad \text{for} \quad j = 1,2, \tag{11.30}$$

where n_i represent the normal vector components to the hole's surface. The role of $\mathbf{U_1}$ is to ensure Neumann boundary conditions hold and the tensor contains simple averages of inverse permittivity and permeability supplemented by the correction term which takes into account the boundary conditions at ∂S_2. Equation (11.28) is the classical expression for the homogenised coefficient in a scalar wave equation with constant material properties; (11.29) is the well-known annex problem of electrostatic type set on a periodic cell, see [4, 11], and also holds for the homogenised vector Maxwell's system, where $\mathbf{U_1}$ now has three components and $i, j = 1,2,3$ [41, 42, 43].

11.2.2.1 Cloaking in metamaterials

SRRs with 4 holes are now used and the dispersion diagrams are in Fig. 11.8 (b). The flat band along the $M\Gamma$ path is interesting for the fifth mode and we choose to illustrate cloaking effects that occur here. In Fig. 11.9(a), we set an harmonic source at the corresponding frequency $\Omega = 2.8$ in an 8×8 array of SRRs and observe a wave pattern of concentric spherical modes. As can be seen in Figs. 11.9(b) and 11.9(c) a plane wave propagating at frequency $\Omega = 2.8$ demonstrates perfect transmission through a slab composed of 38 SRRs but also cloaking of a rectangular inclusion where no scattering is seen before or after the metamaterial slab. Panel (d) of Fig. 11.9 shows the location in the band structure that is responsible for this effect. Note that the frequency of excitation is just below the Dirac cone point located at $\Omega = 2.835$ where the group velocity is negative but also constant near that location of the Brillouin zone illustrated through an isofrequency plot of lower mode of the Dirac point in Fig. 11.9(e). In constrast with the isotropic features of panel (e), those of panels (f) and (g) show ultra-flattened isofrequency contours that relate to ultra-refraction, a regime more prone to omni-directivity than cloaking. The asymptotic system of equations (11.20) describing the effective medium at the Dirac point can be uncoupled to yield one same equation for all $f_0^{(j)}$'s,

$$f^{(j)}_{0,X_iX_i} + 0.7191\Omega_1^4 f_0^{(3)} = 0 \tag{11.31}$$

After some further analysis, the PDE for $f_0^{(2)}$ is responsible for the effects at the frequency chosen $\Omega = 2.8$.

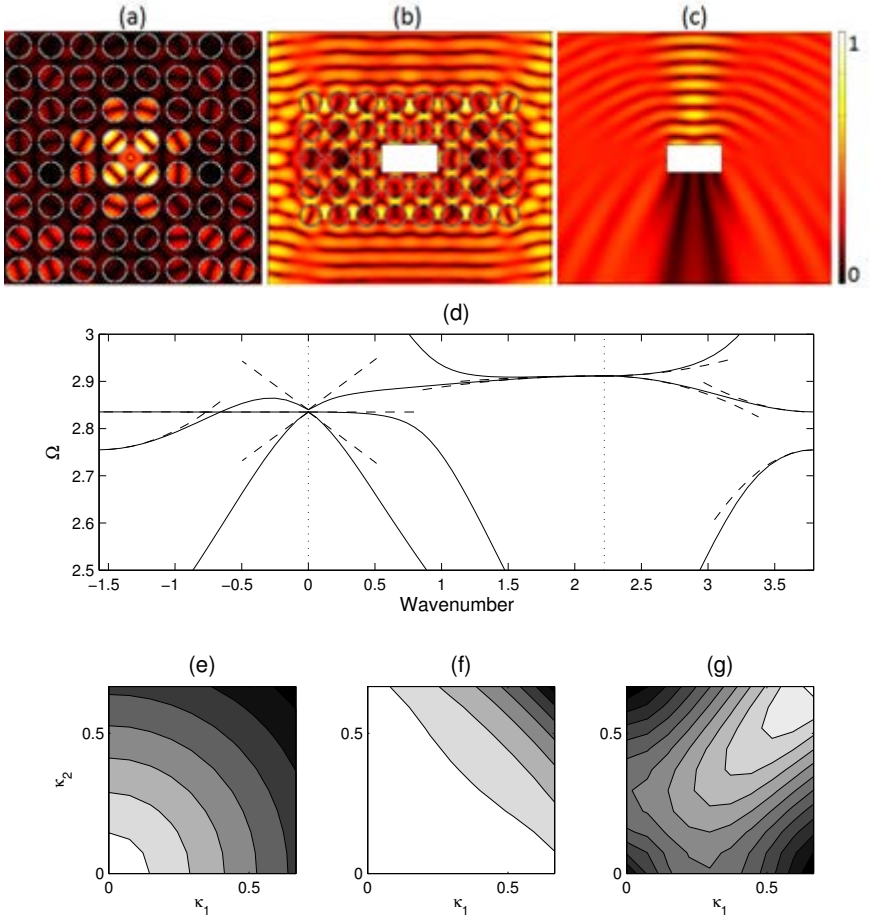


Figure 11.9: Cloaking in square arrays of SRRs with four holes: A source at frequency $\Omega = 2.8$, located in the center of a square metamaterial consisting of 64 SRRs shaped as in Fig. 11.8(b) produces a wave pattern reminiscent of (a) concentric spherical field, (b) cloaking of a rectangular inclusion inside a slab of a metamaterial consisting of 38 SRRs and (c) scattering of a plane wave from the same rectangular hole as the previous panel. (d) Zoom in dispersion diagram of Fig. 11.8(b). Panels (e), (f) and (g) present isofrequency plots of the respective the lower, middle and upper modes of the Dirac point. Figure reproduced from Proceedings of the Royal Society [40].

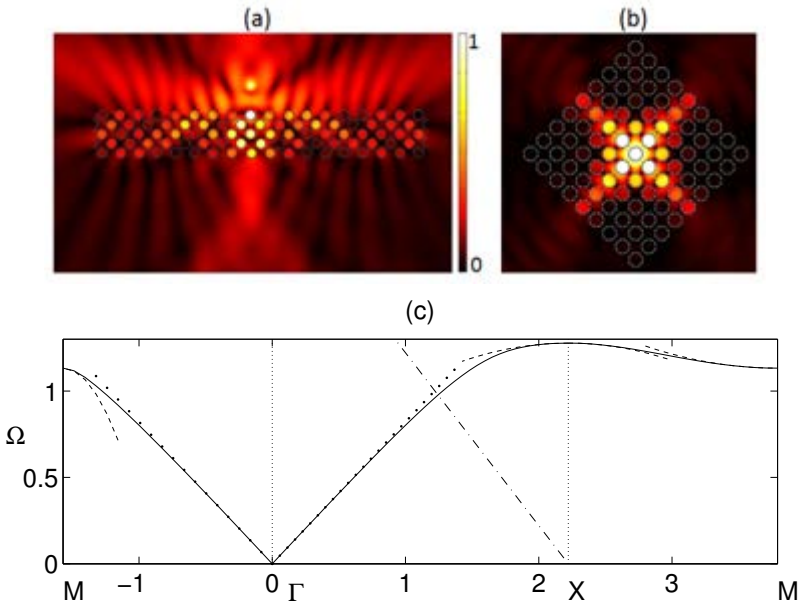


Figure 11.10: Lensing via AANR and St Andrew's cross in square arrays of SRRs with eight holes: (a) A line source at frequency $\Omega = 1.1375$ located above a rectangular metamaterial consisting of of 90 SRRs as in Fig. 11.8(c) displays an image underneath (lensing); (b) A line source at frequency $\Omega = 1.25$ located inside a square metamaterial consisting of 49 SRRs as in Fig. 11.8(c) displays the dynamically induced anisotropy of the effective medium; (c) Zoom in dispersion diagram of Fig. 11.8(c). Note that each cell in the arrays in (a) and (b) has been rotated through an angle $\pi/4$. Figure reproduced from Proceedings of the Royal Society [40].

11.2.2.2 Lensing via AANR and St Andrew's cross in metamaterials

We observe all-angle-negative-refraction effect in metamaterials with SRRs with 8 holes. The dispersion curves in Fig. 11.8(c) are interesting, as the second curve displays the hallmark of an optical band for a photonic crystal (it has a negative group velocity around the Γ point). However, this band is the upper edge of a low frequency stop band induced by the resonance of a SRR, whereas the optical band of a PC results from multiple scattering, which thus arises at higher frequencies. We are therefore in presence of a periodic structure behaving somewhat as a composite intermediate between a metamaterial and a photonic crystal. One of the most topical subjects in photonics is the so-called all-angle-negative- refraction (AANR), which was first described in [46]. AANR allows one to focus light emitted by a point, onto an image, even through a flat lens, provided that certain conditions for AANR are met, such as convex isofrequency contours shrinking with frequency about a point in the Brillouin zone [49].

In Fig. 11.10, we show such an effect for a perfectly conducting photonic crystal (PC) in Fig. 11.10(a). In order to achieve AANR, we choose a frequency on the first dispersion curve (acoustic band) in Fig. 11.8(c), and we take its intersection with the light line $\Omega = \mid \kappa \mid$ along the $X\Gamma$ path. This means that we achieve negative group velocity for waves propagating along the $X\Gamma$ direction of the array, hence the rotation by an angle $\pi/4$ of every cell within the PC in panel (b) of Fig. 11.10. This is a standard trick in optics that has the effect of moving the origin of the light-line dispersion to X as, relative to the PC, the Bloch wavenumber is along $X\Gamma$. This then creates optical effects due to the interaction of the light-line with the acoustic branch, this would be absent if Γ were the light-line origin.

The anisotropy of the effective material is reflected from coefficients $T_{11} = -5.53$ and $T_{22} = 0.2946$. The same frequency of the first band is reachable at point N of the Brillouin zone. By symmetry of the crystal, we would have $T_{11} = 0.2946$ and $T_{22} = -5.53$. The resultant

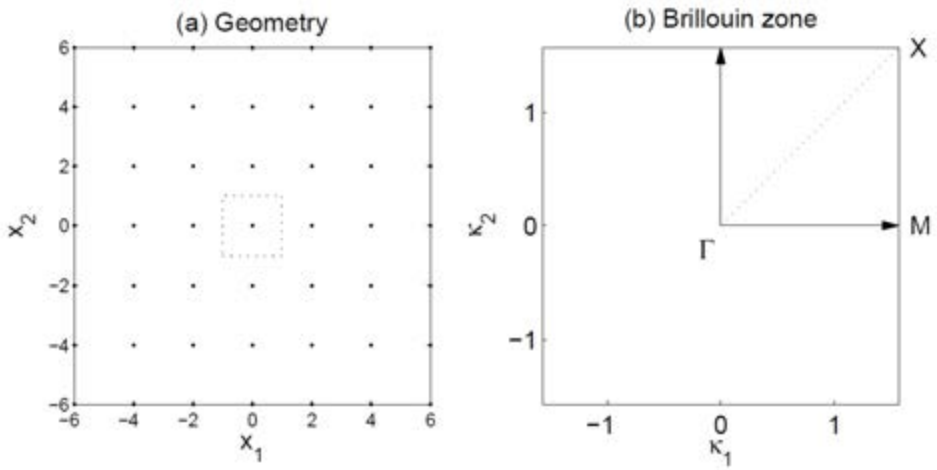


Figure 11.11: For the two dimensional example we show the geometry of the doubly periodic simply supported plate (the dots represent the simple supports) in panel (a) with the elementary cell shown by the dotted lines and in (b) the irreducible Brillouin zone with the lettering for wavenumber positions shown. Figure reproduced from Proceedings of the Royal Society [59].

propagating waves would come from the superposition of the two effective media described above. Fig. 11.10(b) illustrates this anisotropy as the source wave only propagates at the prescribed directions.

11.2.3 Kirchoff Love Plates

HFH is by no means limited to the Helmholtz operator. HFH is here applied to flexural waves in two dimensions [59] for which the governing equation is a fourth order equation

$$\nabla^4 u - \Omega^2 u = 0; \tag{11.32}$$

assuming constant material parameters. Such a thin plate can be subject to point, or line, constraints and these are common place in structural engineering.

In two dimensions, only a few examples of constrained plates are available in the literature: a grillage of line constraints as in [60] that is effectively two coupled one dimensional problems, a periodic line array of point supports [61] raises the possibility of Rayleigh-Bloch modes and for doubly periodic point supports there are exact solutions by [62] (simply supported points) and by [63] (clamped points); the simply supported case is accessible via Fourier series and we choose this as an illustrative example that is of interest in its own right; it is shown in figure 11.11(a). In particular the simply supported plate has a zero-frequency stop-band and a non-trivial dispersion diagram. It is worth noting that classical homogenization is of no use in this setting with a zero frequency stop band. Naturally waves passing through periodically constrained plates have many similarities with those of photonics in optics.

We consider a double periodic array of points at $x_1 = 2n_1, x_2 = 2n_2$ where $u = 0$ (with the first and second derivatives continuous) and so the elementary cell is one in $|x_1| < 1, |x_2| < 1$ with $u = 0$ at the origin (see Figure 11.11); Floquet-Bloch conditions are applied at the edges of the cell.

Applying Bloch's theorem and Fourier series the displacement is readily found [62] as

$$u(\mathbf{x}) = \exp(i\boldsymbol{\kappa}\cdot\mathbf{x}) \sum_{n_1,n_2} \frac{\exp(-i\pi\mathbf{N}\cdot\mathbf{x})}{[(\kappa_1 - \pi n_1)^2 + (\kappa_2 - \pi n_2)^2]^2 - \Omega^2}, \tag{11.33}$$

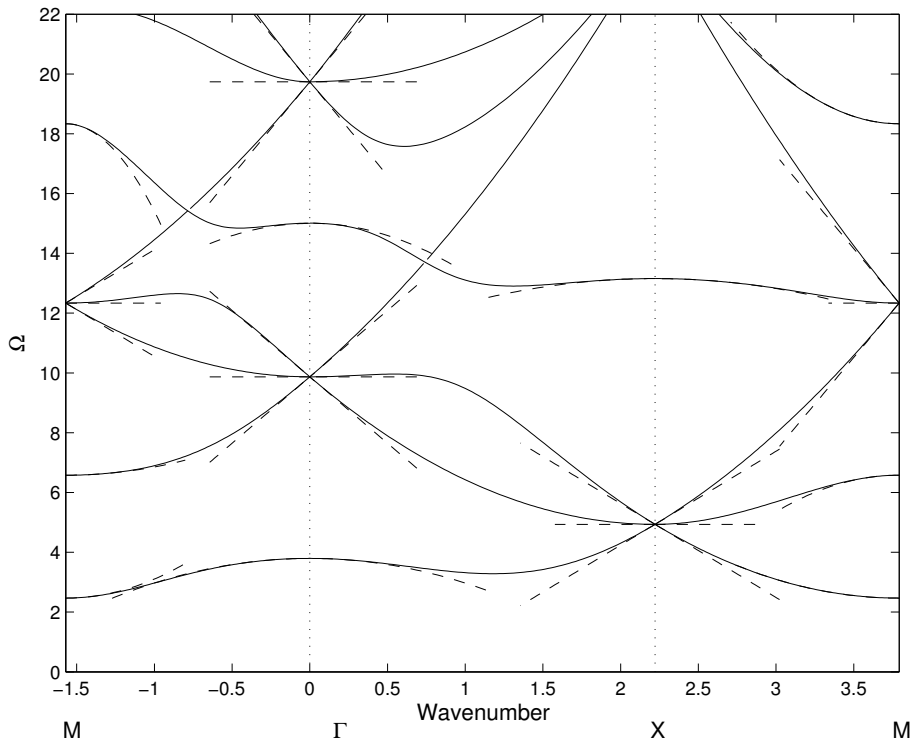


Figure 11.12: The dispersion diagram for a doubly periodic array of point simple supports shown for the irreducible Brillouin zone of Fig. 11.11. The figure shows the dispersion curves as solid lines. As dashed lines, the asymptotic solutions from the high frequency homogenization theory are shown. Figure reproduced from Proceedings of the Royal Society [59]

where $\mathbf{N} = (n_1, n_2)$, and enforcing the condition at the origin gives the dispersion relation

$$D(\kappa_1, \kappa_2, \Omega) = \sum_{n_1,n_2} \frac{1}{[(\pi n_1 - \kappa_1)^2 + (\pi n_2 - \kappa_2)^2]^2 - \Omega^2} = 0, \tag{11.34}$$

In this two dimensional example a Bloch wavenumber vector $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)$ is used and the dispersion relation can be characterised completely by considering the irreducible Brillouin zone ΓXM shown in figure 11.11.

The dispersion diagram is shown in figure 11.12; The singularities of the summand in equation (11.34) correspond to solutions within the cell satisfying the Bloch conditions at the edges, in some cases these singular solutions also satisfy the conditions at the support and are therefore true solutions to the problem, a similar situation occurs in the clamped case considered using multipoles in [63]. Solid lines in figure 11.12 label curves that are branches of the dispersion relation, notable features are the zero-frequency stop-band and also crossings of branches at the edges of the Brillouin zone. Branches of the dispersion relation that touch the edges of the Brillouin zone singly fall into two categories, those with multiple modes emerging at a same standing wave frequency (such as the lowest branch touching the left handside of the figure at M) and those that are completely alone (such as the second lowest branch on the left at M).

The HFH theory can again be employed to find an effective PDE entirely upon the long-scale that describes the behaviour local to the standing wave frequencies and the details are in [59], the asymptotics from the effective PDE are shown in Fig. 11.12 as the dashed lines.

11.3 High-contrast homogenization

Periodic media offer a convenient tool in achieving control of electromagnetic waves, due to their relative simplicity from the point of view of the manufacturing process, and due to the possibility of using the Floquet-Bloch decomposition for the analysis of the spectrum of the wave equation in such media. The latter issue has received a considerable amount of interest in the mathematical community, in particular from the perspective of the inverse problem: how to achieve a given spectrum and/or density of states for the wave operator with periodic coefficients by designing an appropriate periodic structure? While the Floquet-Bloch decomposition provides a transparent procedure for answering the direct question, it does not yield a straightforward way of addressing the inverse question posed above.

One possibility for circumventing the difficulties associated with the inverse problem is by viewing the given periodic structure as a high-contrast one, if this is possible under the values of the material parameters used. The idea of considering high-contrast composites within the context of homogenization appeared first in the work by Allaire [16], which discussed the application of the two-scale convergence technique (Nguetseng [8]) to classical homogenization. A more detailed analysis of high-contrast composites, along with the derivation of an explicit formula for the related spectrum, was carried out in a major study by Zhikov [18]. One of the obvious advantages in using high-contrast composites, or viewing a given composite as a high-contrast one, is in the mere existence of such formula for the spectrum. In the present section we focus on the results of the analysis of Zhikov, and on some more recent results for one-dimensional, layered, high-contrast periodic structures.

In order to get an as short as possible approach to the high-contrast theory, we consider the equation of electromagnetic wave propagation in the transverse electric (TE) polarisation, when the magnetic field has the form $(0,0,H)$, in the presence of sources with spatial density $f(\mathbf{x})$:

$$-\mathrm{div}(\varepsilon^\eta)^{-1}(\mathbf{x}/\eta)\nabla H(\mathbf{x}) = \omega^2 H(\mathbf{x}) + f(\mathbf{x}), \quad \mathbf{x}\in\Omega\subset\mathbb{R}^2, \tag{11.35}$$

where we normalise the speed of light c to 1 for simplicity, which amounts to taking $\varepsilon_0\mu_0 = 1$ in section 11.3, and where the magnetic permeability is assumed to be equal to unity throughout the medium (*i.e.* $\mu = \mu_0$), and the function $f(\mathbf{x})$ is assumed to vanish outside some set that has positive distance to the boundary of Ω. The inverse dielectric permittivity tensor $(\varepsilon^\eta)^{-1}(\mathbf{y})$ is assumed in this section, for simplicity, to be a scalar, taking values $\eta^\gamma I$ and I, respectively, on $[0,1]^2$-periodic open sets F_0 and F_1, such that $\overline{F_0}\cup\overline{F_1} = \mathbb{R}^2$. Here γ is a positive exponent representing a "contrast" between material properties of the two components of the structure that occupy the regions F_0 and F_1. In what follows we also assume that $F_0\cap[0,1]^2$ has a finite distance to the boundary of the unit cell $[0,1]^2$, so that the "soft" component F_0 consists of disjoint "inclusions", spaced $[0,1]^2$-periodically from each other, while the "stiff" component F_1 is a connected subset of $\mathbb{R}^2$. The matrix ε^η represents the dielectric permittivity of the medium at a given point, however the analysis and conclusions of this section are equally applicable to acoustic wave propagation, which is the context we borrow the terms "soft" and "stiff" from. The assumed relation between the values of dielectric permittivity ε^η (in acoustics, between the "stiffnesses") on the two components of the structure is close to the setting of what has been described as "arrow fibres" in the physics literature on electromagnetics, see *e.g* [64].

A simple dimensional analysis shows that if $\omega\sim 1$ then the soft inclusions are in resonance with the overall field if and only if $\gamma = 2$, which is the case we focus on henceforth.

The above equation (11.35) describes the wave profile for a TE-wave in the cylindrical

domain $\Omega \times \mathbb{R}$ domain, and it is therefore supplied with the Neumann condition $\partial H/\partial n = 0$ [3] on the boundary of the domain and with the Sommerfeld radiation condition $\partial H/\partial |x| - \mathrm{i}\omega H = o(|x|^{-1})$ as $|x| \to \infty$.

In line with the previous sections, we apply the method of two-scale asymptotic expansions to the above problem, seeking the solution $H = H(x_1, x_2) = H(\mathbf{x})$ in the form (see also (11.2 in Section 11.1.2)

$$H(\mathbf{x}) = H_0(\mathbf{x}, \mathbf{x}/\eta) + \eta H_1(\mathbf{x}, \mathbf{x}/\eta) + \eta^2 H_2(\mathbf{x}, \mathbf{x}/\eta) + ..., \tag{11.36}$$

where the functions involved are $[0,1]^2$-periodic with respect to the "fast" variable $y = x/\eta$. Substituting the expansion (11.36) into the equation (11.35) and rearranging the terms in the resulting expression in such a way that terms with equal powers of η are grouped together, we obtain a sequence of recurrence relations for the functions H_k, $k = 0, 1, ...$, from which they are obtained sequentially. The first three of these equations can be transformed to the following system of equations for the leading-order term $H^{(0)}(\mathbf{x}, \mathbf{y}) = u(\mathbf{x}) + v(\mathbf{x}, \mathbf{y})$, $\mathbf{x} \in \Omega$, $\mathbf{y} \in [0,1]^2$:

$$-\mathrm{div}\varepsilon_{\mathrm{hom}}^{-1} \nabla u(\mathbf{x}) = \omega^2 \left(u(\mathbf{x}) + \int_{F_0 \cap [0,1]^2} v(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) + f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{11.37}$$

$$-\Delta_{\mathbf{y}} v(\mathbf{x}, \mathbf{y}) = \omega^2 \big(u(\mathbf{x}) + v(\mathbf{x}, \mathbf{y}) \big) + f(\mathbf{x}), \quad y \in F_0 \cap [0,1]^2, \quad v(\mathbf{x}, \mathbf{y}) = 0, \quad y \in F_1 \cap [0,1]^2. \tag{11.38}$$

These equations are supplemented by the boundary conditions for the function u, of the same kind as in the problems with finite η. For the sake of simplifying the analysis, we assume that those inclusions that overlap with the boundary of Ω are substituted by the "main", "stiff" material, where $(\varepsilon^{\eta})^{-1} = I$.

In the equation (11.37), the matrix $\varepsilon_{\mathrm{hom}}$ is the classical homogenization matrix for the perforated medium εF_1, see Section above. However, the properties of the system (11.37)–(11.38) are rather different to those for the perforated-medium homogenised limit, described by the equation $-\mathrm{div}\varepsilon_{\mathrm{hom}}^{-1} \nabla u(\mathbf{x}) = \omega^2 u(\mathbf{x}) + f(\mathbf{x})$. As we shall see next, the two-scale structure of (11.37)–(11.38) means that the description of the spectra of the problems (11.35) in the limit as $\eta \to 0$ diverges dramatically from the usual moderate-contrast scenario.

The true value of the above limiting procedure is revealed by the statement of the convergence, as $\eta \to 0$, of the spectra of the original problems to the spectrum of the limit problem described above, see [18] and by observing that the spectrum of the system (11.37)–(11.37) is evaluated easily as follows. We write an eigenfunction expansion for $v(\mathbf{x}, \mathbf{y})$ as a function of $y \in F_0 \cap [0,1]^2$:

$$v(\mathbf{x}, \mathbf{y}) = \sum_{k=0}^{\infty} c_k(\mathbf{x}) \psi_k(\mathbf{y}), \tag{11.39}$$

where ψ_k are the (real-valued) eigenfunctions of the Dirichlet problem $-\Delta \psi_k = \lambda_k \psi_k$, $y \in F_1 \cap [0,1]^2$, arranged in the order of increasing eigenvalues λ_k, $k = 0, 1, ...$ and orthonormalised according to the conditions $\int_{F_0 \cap [0,1]^2} |\psi_k(\mathbf{y})|^2 d\mathbf{y} = 1$, $k = 0, 1, ...$, and $\int_{F_0 \cap [0,1]^2} \psi_k(\mathbf{y}) \psi_l(\mathbf{y}) d\mathbf{y} = 0$,

[3] Neumann boundary conditions *i.e.* infinite conducting walls is a good model for metals in microwaves, but much less so in the visible range of frequencies wherein absorption by metals need be taken into account. Note also that in the TM polarization case, when the electric field takes the form $(0, 0, E)$, our analysis applies *mutatis mutandis* by interchanging the roles of ε and μ, H and E, and Neumann boundary conditions by Dirichlet ones.

$k \neq l$, $k,l = 0,1,...$ Substituting (11.39) into (11.38), we find the values for the coefficients c_k, which yield an explicit expression for $v(\mathbf{x},\mathbf{y})$ in terms of the function $u(\mathbf{x})$:

$$v(\mathbf{x},\mathbf{y}) = \big(\omega^2 u(\mathbf{x}) + f(\mathbf{x})\big) \sum_{k=0}^{\infty} \Big(\int_{F_0 \cap [0,1]^2} \psi_k(\mathbf{y}) d\mathbf{y} \Big) (\lambda_k - \omega^2)^{-1} \psi_k(\mathbf{y}).$$

Finally, using the last expression in the first equation in (11.37) yields an equation for the function u only:

$$-\mathrm{div} \varepsilon_{\mathrm{hom}}^{-1} \nabla u(\mathbf{x}) = \beta(\omega^2)\big(u(\mathbf{x}) + \omega^{-2} f(\mathbf{x})\big), \quad \mathbf{x} \in \Omega, \tag{11.40}$$

where the function β, which first appeared in the work [18], is given by

$$\beta(\omega^2) = \omega^2 \left(1 + \omega^2 \sum_{k=0}^{\infty} \Big(\int_{F_0 \cap [0,1]^2} \psi_k(\mathbf{y}) d\mathbf{y} \Big)^2 (\lambda_k - \omega^2)^{-1} \right). \tag{11.41}$$

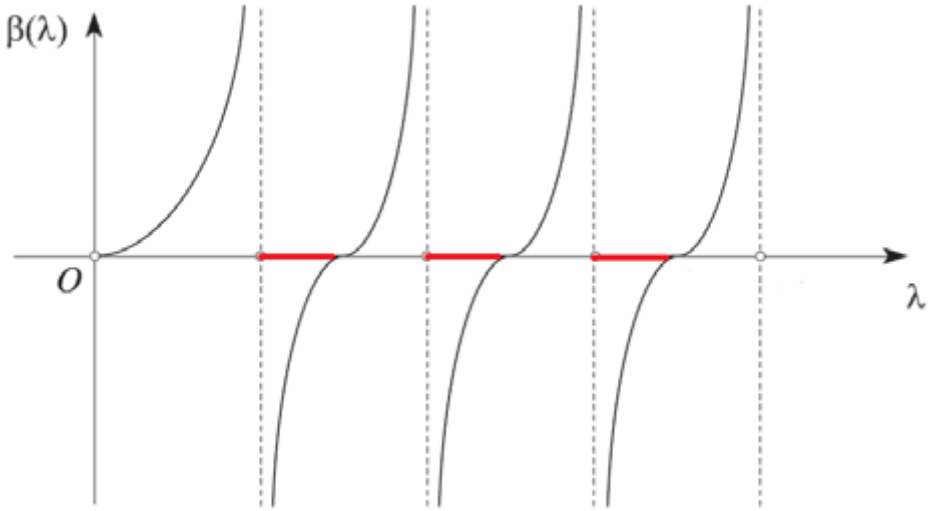


Figure 11.13: The plot of the function β describing the spectrum of the problem (11.37)–(11.38) subject to the boundary conditions. The stop bands for the problem in the whole space $\mathbb{R}^2$ are indicated by the red intervals of the horizontal axis. The spectra of the problems (11.35) considered in the whole space converge, as $\eta \to 0$, to the closure of the complement of the union of the red intervals in the positive semiaxis.

The equation (11.40) is supplemented by appropriate boundary conditions and/or conditions at infinity, which are inherited from the η-dependent family, *i.e.* the Neumann condition at the boundary points $\mathbf{x} \in \partial\Omega$ and the radiation condition when $|\mathbf{x}| \to \infty$. Clearly, the spectrum of this limit problem consists of those values of ω^2 for which $\beta(\omega^2)$ is in the spectrum of the operator generated by the differential expression $-\mathrm{div}\varepsilon_{\mathrm{hom}}^{-1}\nabla$ subject to the same boundary

conditions. For example, for the problem in the whole space $\mathbb{R}^2$ (describing the behaviour of TE-waves in a 3D periodic structure that is invariant in one specified direction) this procedure results in a band-gap spectrum shown in Fig. 11.13. The end points of each pass band are found by a simple analysis of the formula (11.41): the right ends of each pass band are given by those eigenvalues λ_k of the Dirichlet Laplacian on the inclusion $F_0 \cap [0,1]^2$ that possess at least one eigenfunction with non-zero integral over $F_0 \cap [0,1]^2$ (otherwise the corresponding term in (11.41) vanishes), while the left ends of the pass bands are given by solutions to the polynomial equation of infinite order $\beta(\omega^2) = 0$. These points have a physical interpretation as eigenvalues of the so-called electrostatic problem on the inclusion, see [23].

As in the case of classical, moderate-contrast, periodic media, the fact of spectral convergence offers significant computational advantages over tackling the equations (11.35) directly: as $\eta \to 0$ the latter becomes increasingly demanding, while the former requires a single numerical procedure that serves all η once the homogenised matrix ε_{hom} and several eigenvalues λ_k are calculated. A significant new feature, however, as compared to the classical case, is the fact of an infinite set of stop bands opening in the limit as $\eta \to$, which are easily controlled by the explicit description of the band endpoints. This immediately yields a host of applications of the above results for the design of band-gap devices with prescribed behaviour in the frequency interval of interest.

The theorem on spectral convergence for problems described by the equation (11.35) is proved in [18] under the assumption of connectedness of the domain F_1 occupied by the "stiff" component, via a variant of the extension procedure from F_1 to the whole of $\mathbb{R}^2$ for function sequences whose energy scales as η^{-2} (or, equivalently, finite-energy sequences for the operator prior to the rescaling $\mathbf{x}/\eta = \mathbf{y}$). In the more recent works [24], [25], this assumption is dropped in a theorem about spectral convergence for a general class of high-contrast operators, via a version of the two-scale asymptotic analysis akin to (11.36), for the Floquet-Bloch components of the resolvent of the original family of operators following the re-scaling $\mathbf{x}/\eta = \mathbf{y}$. In particular, in [24] a one-dimensional high-contrast model is analysed, which in 3D corresponds to a stack of dielectric layers aligned perpendicular to the direction of the magnetic field. Here the procedure described above for the 2D grating fails to yield a satisfactory limit description as $\eta \to 0$, *i.e.* a description where the spectra of problems for finite η converge to the spectrum of the limit problem described by the system (11.37)–(11.38) as $\eta \to 0$. A more refined analysis of the structure of the related η-dependent family results in a statement of convergence to the set described by the inequalities

$$-1 \leq \frac{1}{2}(\alpha - \beta + 1)\sqrt{\lambda}\sin\Big(\sqrt{\lambda}(\alpha-\beta)\Big) + \cos\Big(\sqrt{\lambda}(\alpha-\beta)\Big) \leq 1. \tag{11.42}$$

where α and β denote the end-points of the inclusion in the unit cell, *i.e.* $F_0 \cap [0,1]^2 = (\alpha,\beta) \times [0,1]$.

Similarly to the spectrum of the 2D high-contrast problem, described by the function β, the limit spectrum of the 1D problem has a band-gap structure, shown in Fig. 11.14, however the description of the location of the bands is different in that it is no longer obtained from the inequality $\beta > 0$, where β is the 1D analogue of (11.41). Importantly, the asymptotic behaviour of the density of states function as $\eta \to 0$ is also very different in the two cases. One can show that the family of resolvents for the problems (11.35) converges, up to a suitable unitary transformation, to the resolvent of a certain operator whose spectrum is given exactly by (11.42), see [25]. The rate of convergence is rigorously shown to be $O(\eta)$, as is anticipated by the expansion (11.36).

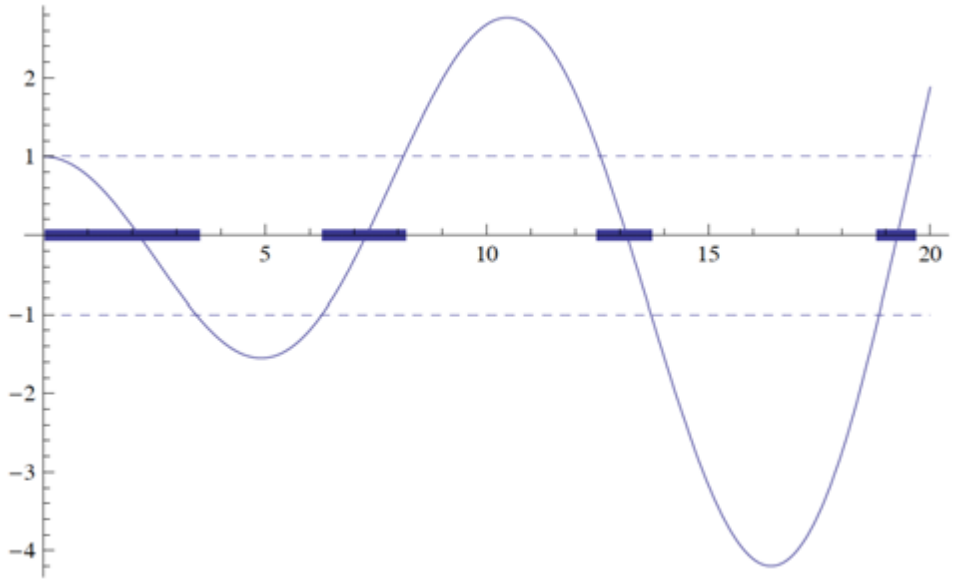


Figure 11.14: The square root of the limit spectrum for a 1D high-contrast periodic stack, in TE polarisation. The oscillating solid line is the graph of the function $f(\omega) = \cos(\omega/2) - \omega\sin(\omega/2)/4$ *in (11.42) with* $\alpha = 1/4$, $\beta = 3/4$. *The square root of the spectrum is the union of the intervals indicated by bold lines.*

The above 1D result is generalised to the case of an oblique incidence of an electromagnetic wave on the same 3D layered structure. Suppose that x_2 is the coordinate across the stack. Then, assuming for simplicity that the wave vector $(\varkappa, 0, 0)$ is parallel to the direction x_1, it can be shown that all three components of the magnetic field are non-vanishing, with the magnetic component $H = H_3$ satisfying the equation

$$-\Big((\varepsilon^\eta)^{-1}(x/\eta)H'(x)\Big)' = \Big(\omega^2 - (\varepsilon^\eta)^{-1}(x/\eta)\varkappa^2\Big)H(x),$$

subject to the same boundary conditions as before. The modified limit spectrum for this family is given by those ω^2 for which (*cf.* (11.42))

$$-1 \le \frac{1}{2}(\alpha-\beta+1)\left(\omega - \frac{\varkappa^2}{\omega}\right)\sin\Big(\sqrt{\lambda}(\alpha-\beta)\Big) + \cos\Big(\sqrt{\lambda}(\alpha-\beta)\Big) \le 1, \quad \omega > 0, \quad (11.43)$$

where, as before, α and β describe the "soft" inclusion layer in the unit cell, see [24]. The set of ω described by the inequalities (11.43) is similar to that shown in Figure 11.14, the only significant difference between the two cases being a low-frequency gap opening near $\omega = 0$ for (11.43).

11.4 Conclusion and further applications to grating theory

To conclude this chapter, we would like to stress that advances in homogenization theory over the past forty years have been fuelled by research in composites [36]. The philosophy of the ne-

cessity for rigour expressed by Lord Rayleigh in 1892 concerning the Lorentz-Lorenz equations (also known as Maxwell-Garnett formulae) can be viewed as the foundation act of homogenization: 'In the application of our results to the electric theory of light we contemplate a medium interrupted by spherical, or cylindrical, obstacles, whose inductive capacity is different from that of the undisturbed medium. On the other hand, the magnetic constant is supposed to retain its value unbroken. This being so, the kinetic energy of the electric currents for the same total flux is the same as if there were no obstacles, at least if we regard the wavelength as infinitely great.' In this paper, John William Strutt, the third Lord Rayleigh [29], was able to solve Laplace's equation in two dimensions for rectangular arrays of cylinders, and in three-dimensions for cubic lattices of spheres. The original proof of Lord Rayleigh suffered from a conditionally convergent sum in order to compute the dipolar field in the array. Many authors in the theoretical physics and applied mathematics communities proposed extensions of Rayleigh's method to avoid this drawback. Another limit of Rayleigh's algorithm is that it does not hold when the volume fraction of inclusions increases. So-called multipole methods have been developed in conjunction with lattice sums in order to overcome such obstacles, see *e.g.* [30] for a comprehensive review of these methods. In parallel to these developments, the quasi-static limit for gratings has been the subject of intensive research, one might cite [31] and [32] for important contributions in the 1980s, and [33] for a comprehensive review of the modern theory of gratings, including a close inspection of homogenization limit.

In order to analyse effective properties of gratings away from the quasi-static limit, the theory of high-frequency homogenization seems a natural way forward. We turn our mind to perfect infinite linear arrays or diffraction gratings constructed periodically and we focus our attention on a single elementary strip of material that then repeats; quasi-periodic Floquet-Bloch boundary conditions describe the phase-shift as a wave moves through the material. Rayleigh-Bloch waves are special as they consist of waves that simultaneously decay exponentially in the perpendicular direction away from the array. Dispersion relations then relate the Floquet-Bloch wavenumber, the phase-shift, to frequency. Although the problem is truly two-dimensional, the assumption of exponential decay in the perpendicular renders it quasi-one dimensional with the wavenumber remaining scalar; this contrasts with the theory of high frequency homogenization in doubly periodic structures, see section 11.2, where a vector wavenumber and the Brillouin zone are more natural.

11.4.1 High-frequency homogenization for gratings

This section draws upon results from [77]. We consider Neumann boundary conditions on the lattice or surface. Recall that this is the TE polarisation for a perfectly conducting surface, and that the governing equation is (11.6) on $-\infty < x_1, x_2 < \infty$, Ω is the non-dimensional frequency and u is the out-of-plane displacement in elasticity or the H_3 component of the magnetic field in TE polarisation.

The two scale nature of the problem is incorporated using the small and large length scales to define two new independent coordinates namely $X = x_1/L$, and $(\xi_1, \xi_2) = (x_1, x_2)/l$. Importantly we just rescale the x_1 coordinate onto a long, scale, (11.7) then becomes,

$$[\partial_{\xi_1\xi_1} + 2\varepsilon\partial_{\xi_1 X} + \varepsilon^2\partial_{XX} + \partial_{\xi_2\xi_2} + \Omega^2]u(X, \xi_1, \xi_2) = 0 \tag{11.44}$$

Standing waves, that exponentially decay perpendicular to the surface/ grating, can occur when there are periodic (or anti-periodic) boundary conditions across the elementary strip (in the $\boldsymbol{\xi}$

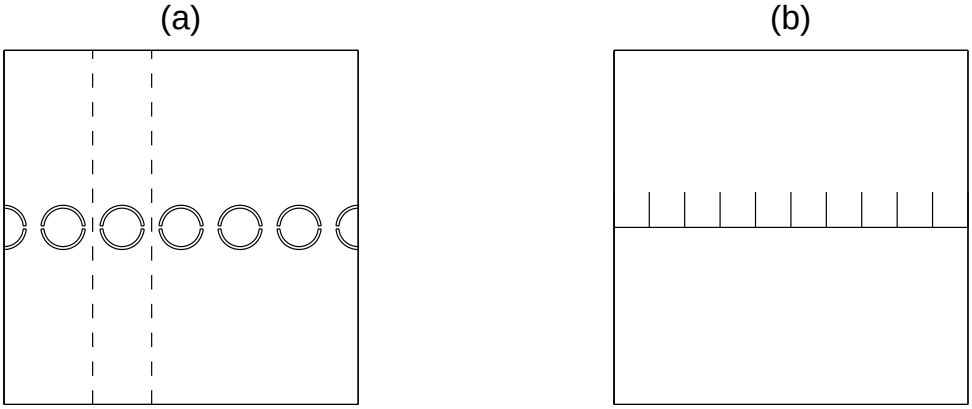


Figure 11.15: Left panel: A diffraction grating of split ring resonators with the elementary cell shown as the dashed strip. Right panel shows a periodic surface that can support spoof surface plasmons [72].

coordinates) and these standing waves encode the local information about the multiple scattering that occurs by the neighbouring strips. The asymptotic technique is then a perturbation about these standing wave solutions, as these are associated with periodic and anti-periodic boundary conditions, which are respectively in-phase and out-of-phase waves across the strip, the conditions in $\boldsymbol{\xi}$ on the edges of the cell, ∂S_1, are known:

$$u|_{\xi_1=1} = \pm u|_{\xi_1=-1} \quad \text{and} \quad u_{,\xi_1}|_{\xi_1=1} = \pm u_{,\xi_1}|_{\xi_1=-1}, \tag{11.45}$$

with the $+,-$ for periodic or anti-periodic cases respectively. Typically, there is only one branch of the dispersion diagram and the periodic case corresponds to long-waves relative to the structure - this case is not particularly interesting and is captured by conventional low-frequency homogenisation. We therefore concentrate upon the anti-periodic case.

We pose the ansatz (11.8) for the field and the frequency and the $u_i(X,\boldsymbol{\xi})$'s adopt the boundary conditions (11.45) on the edge of the cell. An ordered set of equations emerge indexed with their respective power of ε, and are treated in turn

$$u_{0,\xi_i\xi_i} + \Omega_0^2 u_0 = 0, \tag{11.46}$$

$$u_{1,\xi_i\xi_i} + \Omega_0^2 u_1 = -2u_{0,\xi_1 X} - \Omega_1^2 u_0, \tag{11.47}$$

$$u_{2,\xi_i\xi_i} + \Omega_0^2 u_2 = -u_{0,XX} - 2u_{1,\xi_1 X} - \Omega_1^2 u_1 - \Omega_2^2 u_0\ , \tag{11.48}$$

which are the counterparts of (11.46), (11.47) and (11.48). The leading order equation (11.46) is independent of the longscale X and is a standing wave on the elementary cell excited at a specific eigenfrequency Ω_0 and associated eigenmode $U_0(\boldsymbol{\xi};\Omega_0)$, modulated by a long scale function $f_0(X)$ and so we expect to get an ordinary differential equation for f as an effective boundary condition or interface condition characterising the grating when viewed from afar $u_0(X,\boldsymbol{\xi}) = f_0(X)U_0(\boldsymbol{\xi};\Omega_0)$. The entire aim is to arrive at an ODE for f_0 posed entirely upon the longscale, but with the microscale incorporated through coefficients that are integrated, not necessarily averaged, quantities.

Before we continue to next order, equation (11.10), we define the Neumann boundary conditions on the holes or the surface, ∂S_2, as

$$\frac{\partial u}{\partial \mathbf{n}} = u_{,x_i} n_i|_{\partial S_2} = 0. \tag{11.49}$$

where $\mathbf{n}$ is the outward pointing normal, which in terms of the two-scales and $u_i(X,\boldsymbol{\xi})$ become

$$U_{0,\xi_i}n_i = 0, \quad U_0 f_{0,X}n_1 + u_{1,\xi_i}n_i = 0, \quad u_{1,X}n_1 + u_{2,\xi_i}n_i = 0. \tag{11.50}$$

The leading order eigenfunction $U_0(\boldsymbol{\xi};\Omega_0)$ must satisfy the first of these conditions.

Moving to the first order equation (11.47) we invoke a solvability condition by integrating over the cell the product of equation (11.47) and U_0 minus the product of equation (11.46) and $u_1/f_0(\mathbf{X})$. The eigenvalue Ω_1 is zero. We solve for $u_1 = f_{0,X}U_1(\boldsymbol{\xi})$, so U_1 is a scalar this time. So at first order

$$\nabla^2_{\boldsymbol{\xi}}U_1 + \Omega_0^2 U_1 = -2U_{0,\xi_1} \tag{11.51}$$

this is solved subject to

$$\mathbf{n}\cdot\nabla_{\boldsymbol{\xi}}U_1 = -U_0 n_1 \tag{11.52}$$

on the boundary. By invoking a similar solvability condition for equation (11.48) we obtain the desired ordinary differential equation for f_0

$$T f_{0,XX} + \Omega_2^2 f_0 = 0 \tag{11.53}$$

posed entirely on the longscale X. The tensor t_{ij} consists of integrals over the microcell in $\boldsymbol{\xi}$ and is ultimately independent of $\boldsymbol{\xi}$ and in this case is just a scalar T. The formulation for T reads,

$$T\int\int_S U_0^2 dS = \int\int_S (U_0^2 + 2U_{1,\xi_1}U_0)dS - \int_{\partial S_2} U_1 U_0 n_1 ds, \tag{11.54}$$

where by using Green Theorem with vector field $\mathbf{F} = (U_1U_0, 0)$ equation (11.54) simplifies to,

$$T\int\int_S U_0^2 dS = \int\int_S (U_0^2 + U_{1,\xi_1}U_0 - U_{0,\xi_1}U_1)dS, \tag{11.55}$$

Therefore, if the surface or grating supports Rayleigh-Bloch waves then these are represented as an effective string or membrane equation similar to (11.15).

We now illustrate the theory using arrays of circular holes and the comb-like structure.

11.4.2 Illustrations for the classical comb and SRR gratings

An early example for which Rayleigh-Bloch waves were found explicitly is that of a Neumann comb-like surface consisting of periodic thin plates of finite length, a, perpendicular to a flat wall. This was initially studied by Hurd [66] with later modifications by [67, 68, 70]. In particular the analytical approach using the residue calculus method has enabled the question of whether embedded Rayleigh-Bloch waves exist to be studied [70] explicitly.

We will concentrate upon the non-embedded Rayleigh-Bloch waves in $\Omega < \kappa$ and Hurd's dispersion relation

$$\Omega a = (n+1/2)\pi + \Omega/\pi \ln 2 + \chi(\kappa,\Omega) \tag{11.56}$$

where

$$\chi(\kappa,\Omega) = \sum_{n=1}^{\infty}\left(\sin^{-1}(\Omega/n\pi) - \sin^{-1}(\Omega/(\beta+2n\pi)) - \sin^{-1}(\Omega/|\beta-2n\pi|)\right) - \sin^{-1}(\Omega/\kappa) \tag{11.57}$$

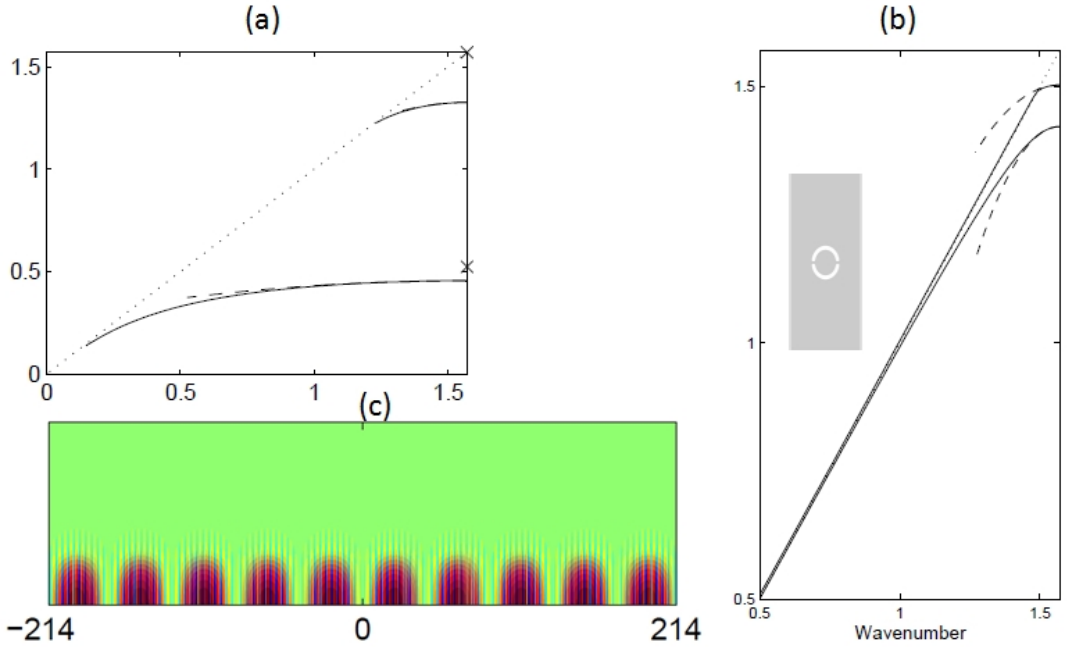


Figure 11.16: (a) The two dispersion branches below cutoff ($\Omega = \kappa$) for the comb-like structure with pins of width 0.05 and height 3. (b) The two dispersion branches below cutoff ($\Omega = \kappa$) for the SRR-like structure. (c) Plot of real part of field u near the lowest cut-off frequency in panel (a) generated by a point source at normalized frequency $\Omega = 0.4509$. Reproduced from Proceedings of the Royal Society [77]

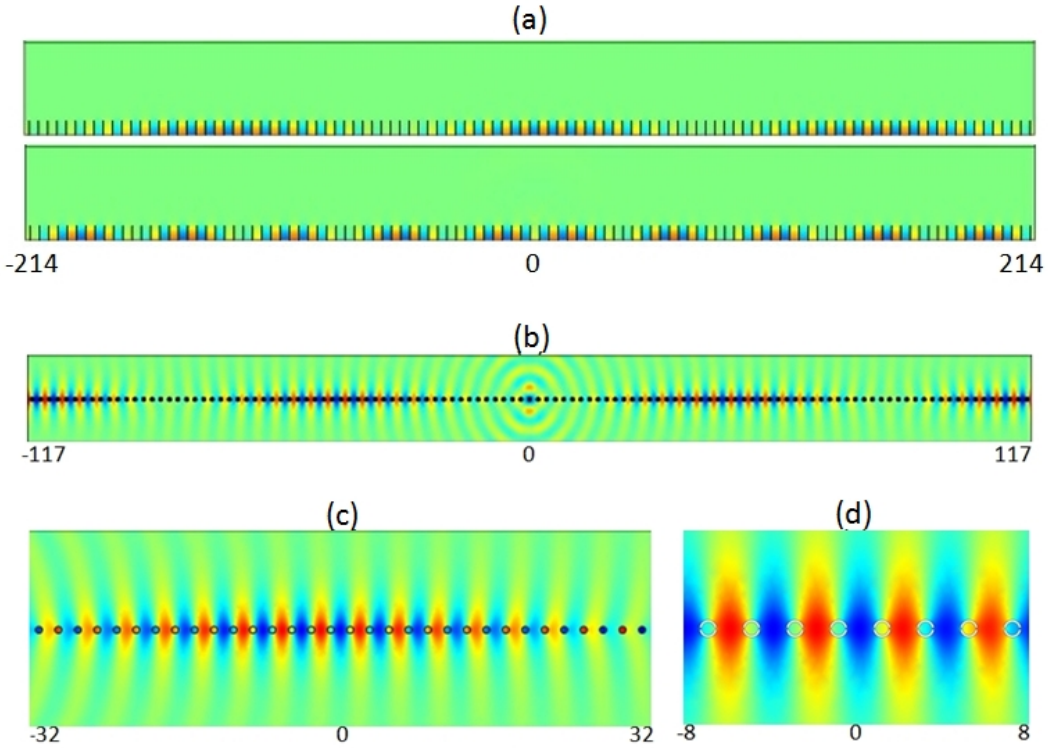


Figure 11.17: Panel (a) Plots of real part of field u for standing and near cut-off frequencies in a comb grating consisting of pins of width 0.05 and height 3: (a) Fields generated by a point source at normalized frequencies $\Omega = 0.4513$ (standing wave) and $\Omega = 0.45$ (near cut-off wave). Panels (b-d) correspond to plots of real part of field u near cut-off frequencies 1.45 in a SRR grating with SRR of inner and outer radii 0.85 and 0.95 and a cut of thickness 0.06. Reproduced from Proceedings of the Royal Society [77]

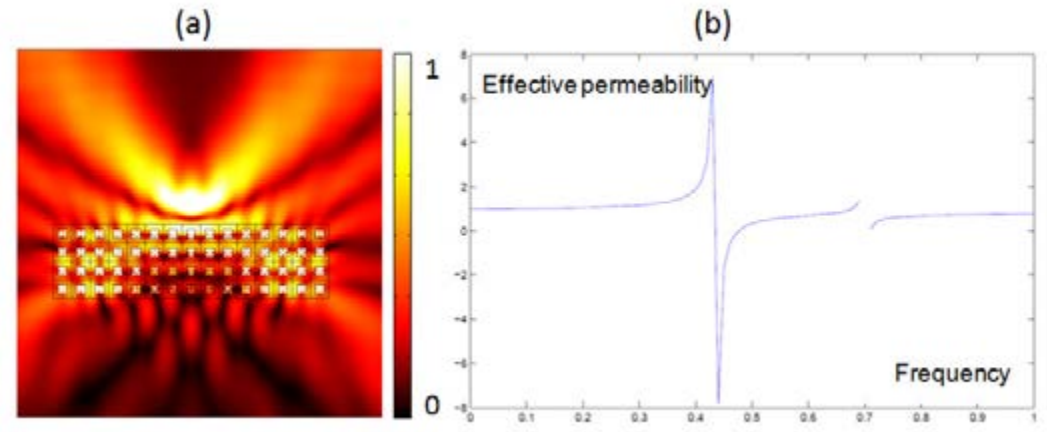


Figure 11.18: Superlens application of grating: (a) A time harmonic source at frequency 0.473 *displays an image through a square array of square inclusions; (b) Effective magnetism versus frequency using (11.58) for square inclusions of relative permittivity* 100 *with sidelength* $a = 0.5d$ *in matrix of relative permittivity* 1 *(grating pitch* $d = 0.1$*); Negative values of the effective magnetism are in the frequency region* $[0.432, 0.534]$.

provides a highly accurate approximation; a couple of dispersion branches are shown in Fig. 11.16 using Hurd's formulae, and also a more accurate result from [68] which is virtually indistinguishable from that of Hurd. Here a is the length of the tooth, and the curves are locally quadratic near π as we expect. If we increase a then more dispersion curves occur.

Above $\Omega = \kappa$ there are embedded Rayleigh-Bloch surface waves, as described in [70, 71], indeed [71], figure 11.16 appears to show Rayleigh-Bloch dispersion curves above the second cut-off. An example of surface wave obtained with HFH is shown in panel (c), and should be compared to panel (a) in figure 11.17.

Interestingly, in the pure mathematics community, Zhikov's work on high-contrast homogenization [18] has had important applications in metamaterials, with the interpretation of his homogenized equations in terms of effective magnetism first put forward by O'Brien and Pendry [65], and then by Bouchitté and Felbacq [73], although these authors did not seem to be aware at that time of Zhikov's seminal paper [18]. In order to grasp the physical importance of (11.40)-(11.41), we consider the case of square inclusions of sidelength $a = d/2$, where d is the pitch of a bi-periodic grating. The eigenfunctions are $\psi_{nm}(\mathbf{y}) = 2\sin(n\pi y_1)\sin(n\pi y_2)$ in (11.41) and the corresponding eigenvalues are $k_{nm}^2 = \pi^2(n^2+m^2)$. The right-hand side in the homogenized equation (11.40) can then be interpreted in terms of effective magnetism:

$$\mu_{hom}(k) = 1 + \frac{64a^2}{\pi^4} \sum_{(n,m)odd} \frac{k^2}{n^2m^2(k_{nm}^2/a^2 - k^2)} \,. \tag{11.58}$$

This function can be computed numerically for instance with Matlab and demonstrates that negative values can be achieved for μ_{hom} near resonances, see Fig. 11.18(b). This allows for superlensing via negative refraction, as shown in Fig. 11.18(a).

Finally, we would like to point out that high-order homogenization techniques [74] suggest that most gratings display some artificial magnetism and chirality when the wavelength is no longer much larger than the periodicity [75, 76]. We hope we have convinced the reader that there is a whole new range of physical effects in gratings which classical, high-frequency and high-contrast homogenization theories can capture.

References:

[1] Petit, R., 1980. *Electromagnetic theory of gratings*, Topics in current physics, Springer-Verlag, Berlin.

[2] Bakhvalov, N. S.,1975. Averaging of partial differential equations with rapidly oscillating coefficients. *Dokl. Akad. Nauk SSSR* **221**, 516–519. English translation in *Soviet Math. Dokl.* **16**, 1975.

[3] De Giorgi, E., Spagnolo, S., 1973. Sulla convergenza degli integrali dell'energia per operatori ellitici del secondo ordine. *Boll. Unione Mat. Ital., Ser* **8***,* 391–411.

[4] Bensoussan, A., Lions, J.L., Papanicolaou, G., 1978. *Asymptotic analysis for periodic structures*, North-Holland, Amsterdam

[5] Marchenko, V. A., Khruslov, E. Ja., 1964. Boundary-value problems with fine-grained boundary. (Russian) *Mat. Sb. (N.S.)* **65** (107) 458–472.

[6] Tartar, L., 1974. Problème de côntrole des coefficients dans des équations aux dérivées partielles. *Lecture Notes in Economics and Mathematical Systems* **107**, 420–426.

[7] Murat, F., 1978. Compacité par compensation. (French) *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **5**(3), 489–507.

[8] Nguetseng, G., 1989. A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal.* **20** (3), 608–623.

[9] Cioranescu, D., Damlamian, A., Griso, G., 2002. Periodic unfolding and homogenization, *C. R. Math. Acad. Sci. Paris,* **335**, 99-104.

[10] Bakhvalov, N. S., Panasenko, G. P., 1984. *Homogenization: Averaging Processes in Periodic Media.* Nauka, Moscow (in Russian). English translation in: Mathematics and its Applications (Soviet Series) **36**, Kluwer Academic Publishers.

[11] Jikov, V. V., Kozlov, S. M., Oleinik, O. A., 1994. *Homogenization of Differential Operators and Integral Functionals.* Springer, Berlin.

[12] Sanchez-Palencia, E., 1980 *Nonhomogeneous media and Vibration Theory.* Lecture Notes in Physics **127**, Springer, Berlin.

[13] Chechkin, G. A., Piatnitski, A. L., Shamaev, A. S., 2007. *Homogenization: Methods and Applications.* AMS Translations of Mathematical Monographs **234**.

[14] Kozlov, S. M., 1979. The averaging of random operators. *Mat. Sb. (N.S.),* **109**(2),188-202.

[15] Papanicolaou, G. C., Varadhan, S. R. S., 1981. Boundary value problems with rapidly oscillating random coefficients. Random fields, Vol. I, II (Esztergom, 1979), 835–873, *Colloq. Math. Soc. János Bolyai,* **27**, North-Holland, Amsterdam-New York.

[16] Allaire, G., 1992. Homogenization and two-scale convergence, *SIAM J. Math. Anal.* **23**, 1482–1518.

[17] Arbogast, T., Douglas, J., Hornung, U., 1990. Derivation of the double porosity model of single phase flow via homogenization theory. *SIAM J. Math. Anal.* **21** (4), 823-836.

[18] Zhikov, V. V., 2000. On an extension of the method of two-scale convergence and its applications, *Sb. Math.*, **191**(7), 973–1014.

[19] Kozlov, V., Mazya, V., Movchan, A., 1999. *Asymptotic Analysis of Fields in Multi-structures,* Oxford University Press, Oxford.

[20] Mazya, V. G., Nazarov, S.A., Plamenevskii, B. A., 2000. *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains.* Vol. I, Operator Theory: Advances and Applications **111**, Birkäuser Verlag, Berlin.

[21] Zhikov, V. V. , 2002. Averaging of problems in the theory of elasticity on singular structures. (Russian) *Izv. Ross. Akad. Nauk Ser. Mat.* **66** (2), 81–148. English translation in *Izv. Math.* **66** (2002), No. 2, 299–365.

[22] Zhikov, V. V., Pastukhova, S. E., 2002. Averaging of problems in the theory of elasticity on periodic grids of critical thickness. (Russian) *Dokl. Akad. Nauk* **385** (5), 590–595.

[23] Zhikov, V. V., 2005. On spectrum gaps of some divergent elliptic operators with periodic coefficients. *St. Petersburg Math. J.* **16**(5) (2005), 774–790.

[24] Cherednichenko. K. D., Cooper, S., Guenneau, S., 2012. Spectral analysis of one-dimensional high-contrast elliptic problems with periodic coefficients, *Submitted.*

[25] Cherednichenko, K. D., Cooper, S., 2012. On the resolvent convergence of periodic differential operators with high contrast coefficients, *Preprint.*

[26] Zolla, F., Guenneau, S., 2003. Artificial ferro-magnetic anisotropy : homogenization of 3D finite photonic crystals, *in Movchan Ed. Asymptotics, singularities and homogenization in problems of mechanics*, Kluwer Academic Press, 375–385.

[27] Pendry, J.B., Schurig, D., Smith, D.R., 2006. Controlling Electromagnetic Fields, Science **312**, 1780-1782.

[28] Huang, Y., Feng, Y., Jiang, T., 2007. Electromagnetic cloaking by layered structure of homogeneous isotropic materials, Optics Express **15**(18), 11133-11141.

[29] Strutt, J.W. (Lord Rayleigh) 1892. On the influence of obstacles arranged in rectangular order upon the properties of a medium, Phil. Mag **34**, 481-502.

[30] Movchan, A.B., Movchan, N.V., Poulton, C.G., 2002. *Asymptotic models of fields in dilute and densely packed composites*, Imperial College Press, London.

[31] McPhedran, R.C., Botten, L.C., Craig, M.S., Neviere, M., Maystre, D., 1982. Lossy Lamelar gratings in the quasistatic limit, Optica Acta **29**, 289-312.

[32] Petit, R., Bouchitte, G., 1987. Replacement of a very fine grating by a stratified layer: homogenisation techniques and the multiple-scale method, *SPIE Proceedings, Application and Theory of Periodic Structures, Diffraction Gratings, and Moiré Phenomena* **431**, 815.

[33] Neviere, M., Popov, E., 2003. *Light Propagation in Periodic Media: Diffraction Theory and Design*, Marcel Dekker, New York.

[34] Craster, R.V., Guenneau, S., 2013. *Acoustic Metamaterials: Negative Refraction, Imaging, Lensing and Cloaking*, Springer Series in Materials Science **166**, Springer-Verlag, Berlin.

[35] Mei, C.C., Auriault, J.-L., Ng, C-O., 1996. Some applications of the homogenization theory, Adv. Appl. Mech. **32**, 277-348.

[36] Milton, G. W., 2002. *The Theory of Composites*, Cambridge University Press, Cambridge.

[37] Craster, R.V., Kaplunov, J., Pichugin, A.V., 2010. High frequency homogenization for periodic media, Proc R Soc Lond A **466**, 2341-2362.

[38] Nemat-Nasser, S., Willis, J. R., Srivastava, A., Amirkhizi, A. V., 2011. Homogenization of periodic elastic composites and locally resonant sonic materials, Phys. Rev. B **83**, 104103.

[39] Craster, R.V., Antonakakis, T., Makwana, M., Guenneau, S., 2012. Dangers of using the edges of the Brillouin zone, Phys. Rev. B **86**, 115130.

[40] Antonakakis, T., Craster, R.V., Guenneau, S., 2013. Asymptotics for metamaterials and photonic crystals, Proc R Soc Lond A, (http://dx.doi.org/10.1098/rspa.2012.0533)

[41] Guenneau, S., Zolla, F., 2000. Homogenization of three-dimensional finite photonic crystals, Progress In Electromagnetic Research **27**, 91-127 (http://www.jpier.org/PIER/pier27/9907121jp.Zolla.pdf)

[42] Wellander, N., Kristensson, G. 2003. Homogenization of the Maxwell equations at fixed frequency, SIAM J. Appl. Math. **64**(1), 170–195.

[43] Guenneau, S., Zolla, F., Nicolet, A. 2007. Homogenization of 3D finite photonic crystals with heterogeneous permittivity and permeability, Waves in Random and Complex Media **17**(4), 653–697.

[44] E. Yablonovitch, 1987. Inhibited spontaneous emission in solid-state physics and electronics, Phys. Rev. Lett. **58**, 2059-2062.

[45] S. John, 1987. Strong localization of photons in certain disordered dielectric superlattices, Phys. Rev. Lett. **58**, 2486-2489.

[46] Zengerle, R., 1987. Light propagation in singly and doubly periodic waveguides, J. Mod. Opt. **34**, 1589-1617.

[47] Gralak, B., Enoch, E., Tayeb, G., 2000. Anomalous refractive properties of photonic crystals, J. Opt. Soc. Am. A **17**, 1012-1020.

[48] Notomi, N., 2000. Theory of light propagation in strongly modulated photonic crystals: Refractionlike behaviour in the vicinity of the photonic band gap, Phys. Rev. B **62**, 10696-10705.

[49] Luo, C., Johnson, S.G., Joannopoulos, J.D., 2002. All-angle negative refraction without negative effective index, Phys. Rev. B **65**, 201104(R).

[50] Dowling, J. P., Bowden, C. M., 1994. Anomalous index of refraction in photonic bandgap materials, J. Mod. Optics **41**, 345-351.

[51] Enoch, S., Tayeb, G., Sabouroux, P., Guerin, N., Vincent, P., 2002. A metamaterial for directive emission, Phys. Rev. Lett. **89**, 213902.

[52] Craster, R. V., Kaplunov, J., Nolde, E., Guenneau, S., 2011. High frequency homogenization for checkerboard structures: Defect modes, ultrarefraction and all-angle-negative refraction. J. Opt. Soc. Amer. A **28**, 1032-1041.

[53] Allaire, G., Conca C., 1998. Bloch wave homogenization and spectral asymptotic analysis, J. Math. Pures. Appl. **77**, 153-208.

[54] Allaire, G., Piatnitski, A., 2005. Homogenisation of the Schrödinger equation and effective mass theorems, Commun. Math. Phys. **258**, 1-22.

[55] Birman, M. S., Suslina, T. A., 2006. Homogenization of a multidimensional periodic elliptic operator in a neighborhood of the edge of an internal gap, J. Math. Sciences **136**, 3682-3690.

[56] Hoefer, M. A., Weinstein, M. I., 2011. Defect modes and homogenization of periodic Schrödinger operators, SIAM J. Math. Anal. **43**, 971-996.

[57] Parnell, W. J., Abrahams, I. D., 2006. Dynamic homogenization in periodic fibre reinforced media. Quasi-static limit for SH waves, Wave Motion **43**, 474-498.

[58] Cherednichenko, K., Smyshlyaev, V. P. and Zhikov, V. V., 2006. Non-local homogenised limits for composite media with highly anisotropic periodic fibres, Proc. R. Soc. Ed. A **136**(1), 87-114.

[59] Antonakakis, T., Craster, R.V., 2012. High frequency asymptotics for microstructured thin elastic plates and platonics, Proc R Soc Lond A **468**, 1408-1427.

[60] Mace, B.R., 1981. Sound radiation from fluid loaded orthogonally stiffened plates, J. Sound Vib. **79**, 439-452.

[61] Evans, D.V., Porter, R., 2007. Penetration of flexural waves through a periodically constrained thin elastic plate floating in *vacuo* and floating on water, J. Engng. Math. **58**, 317-337.

[62] Mace, B.R., 1996. The vibration of plates on two-dimensionally periodic point supports, J. Sound Vib. **192**, 629-644.

[63] Movchan, A.B., Movchan, N.V., McPhedran, R.C.,2007. Bloch-Floquet bending waves in perforated thin plates, Proc R Soc Lond A **463**, 2505-2518.

[64] Zolla, F., Renversez, G., Nicolet, A., Kuhlmey, B., Guenneau, S., Felbacq, D., Argyros, A., Leon-Saval, S., 2012. *Foundations of photonic crystal fibres*, Imperial College Press, London.

[65] OBrien, S., Pendry, J.B., 2002. Photonic Band Gap Effects and Magnetic Activity in Dielectric Composites, J. Phys.: Condensed Matter **14**, 4035-4044.

[66] Hurd, R. A. 1954. The propagation of an electromagnetic wave along an infinite corrugated surface, Can. J. Phys. **32**, 727-734.

[67] DeSanto, J. A. 1972. Scattering from a periodic corrugated structure II. Thin comb with hard boundaries, J. Math. Phys. **13**, 336-341.

[68] Evans, D. V., Linton, C. M. 1993. Edge waves along periodic coastlines, Q. Jl Mech. appl. Math **46**, 643-656.

[69] Evans, D. V., Porter, R. 1998. Trapping and near-trapping by arrays of cylinders in waves, J. Engng Math. **35**, 149-179.

[70] Evans, D. V., Porter, R. 2002. On the existence of embedded surface waves along arrays of parallel plates, Q. Jl Mech. appl. Math **55**, 481-494

[71] Porter R. and Evans D. V. 2005. Embedded Rayleigh-Bloch surface waves along periodic rectangular arrays, Wave Motion **43**, 29-50

[72] Pendry, J. B., Martin-Moreno, L., Garcia-Vidal, F. J., 2004. Mimicking surface plasmons with structured surfaces, Science **305**, 847-848.

[73] Bouchitte, G., Felbacq, D., 2004. Homogenization near resonances and artificial magnetism from dielectrics, C. R. Math. Acad. Sci. Paris **339** (5), 377-382.

[74] Cherednichenko, K. and Smyshlyaev, V. P., 2004. On full two-scale expansion of the solutions of nonlinear periodic rapidly oscillating problems and higher-order homogenised variational problems, Arch. Rat. Mech. Anal. **174** (3), 385-442.

[75] Liu, Y., Guenneau, S., Gralak, B., 2013. Causality and passivity properties of effective parameters of electromagnetic multilayered structures, Phys. Rev. B **88**(16), 165104.

[76] Liu, Y., Guenneau, S., Gralak, B., 2013. Artificial dispersion via high-order homogenization : magnetoelectric coupling and magnetism from dielectric layers, Proc. Roy. Soc. Lond. A **469**(2158), 20130240.

[77] Antonakakis T., Craster R. V., Guenneau S., Skelton E. A., 2014. An asymptotic theory for waves guided by diffraction gratings or along microstructured surfaces, Proc. R. Soc. Lond. A **470**, 20130467.

Chapter 12:
Boundary Integral Equation Methods for Conical Diffraction and Short Waves

Leonid I. Goray and Gunther Schmidt

Table of Contents:

Chapter 12

Boundary Integral Equation Methods for Conical Diffraction and Short Waves

Leonid I. Goray[1,2] and Gunther Schmidt[3]

[1] *Saint Petersburg Academic University, RAS, Khlopin 8/3, St. Petersburg 194021, Russian Federation, goray@spbau.ru,*

[2] *Institute for Analytical Instrumentation, RAS, Ivana Chernykh 31-33, St. Petersburg 198095, Russian Federation, lig@pcgrate.com,*

[3] *Weierstrass Institute of Applied Analysis and Stochastics, Mohrenstrasse 39, Berlin 10117, Germany, schmidt@wias-berlin.de*

12.1 Introduction

This work is part of research that has been pursued by the authors over a long period of time for the purpose of developing accurate and fast numerical algorithms, including the commercial packages PCGrate and DiPoG [12.1, 12.2] designed to model multilayered gratings having mostly one-dimensional periodicity (1D), including roughness, and working in all, including the shortest, optical wavelength ranges at arbitrary optical mounts.

The boundary integral equation theory or, briefly, integral method (IM) is presently universally recognized as one of the most developed and flexible approaches to an accurate numerical solution of diffraction grating problems (see, e.g., Ref. 12.3 and Ch. 4 and references therein). Viewed in the historical context, this method was the first to offer a solution to vector problems of light diffraction by optical gratings and to demonstrate remarkable agreement with experimental data. This should be attributed to the high accuracy and good convergence of the method, especially for the TM polarization plane. It does not involve limitations similar to those characteristic of the Coupled-Wave Analysis (CWA), and it provides a better convergence. The disadvantages of this method include its being mathematically complicated, as well as numerous "peculiarities" involved in numerical realization. In particular, quasi-periodic Greens functions and their derivatives appearing as kernels in the integral operators require sophisticated lattice sum techniques to evaluate. Moreover, application of the IM to cases of heterogeneous or anisotropic media meets with difficulties; however, with the volume integral method it is possible to overcome these difficulties easily. Nevertheless, it is on the basis of this theory that all the well-known problems of diffraction by periodic and non-periodic structures in optics and other fields have been solved. In many cases it offers the only possible way to follow up in research. The flexibility and universality inherent in the IM, in particular, enable

one rather easily to reduce the problem of radiation of Gaussian waves or of a localized source to that of plane-wave incidence, for which scientists all over the world have a set of numerical solutions. Generalizations of the IM have recently been proposed for arbitrarily profiled 1D multilayer gratings [12.4], randomly-rough x-ray-extreme-ultraviolet (EUV) gratings and mirrors [12.5, 12.6], conical diffraction gratings including materials with a negative permittivity and permeability (metamaterials) [12.7, 12.8], bi-periodic anisotropic structures using a variation formulation [12.9], Fresnel zone plates and diffraction optical elements [12.10, 12.11], and two-dimensional (2D) [12.12, 12.13] and three-dimensional (3D) [12.14] photonic crystals (inclusions) of some geometries, among others.

The IM is so pivotal that one can indicate the few areas where it can be modified and improved to solve particular diffraction problems. By convention they are: (1) physical model—choice of boundary types, boundary conditions, layer and substrate refractive indices, and radiation conditions; (2) mathematical structure—integral representations using potentials or integral formulas and a multilayer scheme; (3) method of approximation and discretization—discretization schemes, choice of basis (trial) and test (weighting) functions, and treatment of coincident points and corners in boundary profile curves; (4) low-level details—calculations and optimization of kernel functions, mesh of discretization (collocation) points, quadrature rules, and solution of linear algebraic systems; (5) implementation enhancements—memory caching, other implementation details. A self-consistent explanation of the existing IMs is beyond the primary scope of the present study. The main purpose of this Chapter is to present a complete description in general operator form of the two IMs applied to 1D multilayer gratings working in conical diffraction mounts and in short waves. Our study also includes the calculus of grating absorption in the explicit form and scattering intensity of randomly-rough gratings using Monte Carlo simulations. For other formal IM treatments and their comparisons, one should rather look to the references of this Chapter as well as to Ch. 4 and to references therein.

Various kinds of electromagnetic features of different nature can exist and be explored in complex grating structures: Bragg and Brewster resonances, Rayleigh anomalies and groove shape features, waveguiding and Fano-type modes, etc. In conical diffraction, the influence of possible types of waves can be mixed. For the purposes of this Chapter, we chose three important types, among many others, of diffraction grating problems to include them in Section 12.9 "Examples of numerical results". They are: bare dielectric or metallic gratings of standard groove shapes working in conical diffraction in the resonance domain; shallow high-conductive or dielectric gratings of various boundary shapes, including closed ones, working in different mounts and supporting polariton-plasmon excitation or Bragg diffraction in the visible–infrared range; bare and multilayer gratings working in grazing-conical or near-normal in-plane diffraction in the soft x-ray–EUV range.

12.2 Integral method for one-profile gratings in conical diffraction

The present IM designed for the calculation of the efficiency of bulk and multilayer gratings with arbitrary boundary shapes including micro- and nanoroughness and over an extremely wide wavelength range is considered here in a general operator formalism. In this Section, we consider single-boundary integral equations, which involve boundary integrals of the single and double layer potentials and also the normal and tangential derivatives of single layer potentials. Analytical aspects of boundary integral operators are well represented in publications and, the most relevant of them for the present study, are also in following Sections. The fields are assumed time harmonic. Under these conditions, in classical (in-plane) diffraction the Maxwell

system of equations reduces to a single Helmholtz equation; therefore fields are represented in the sequel by scalar functions. They would be vector functions with two components in the case of conical (off-plane) diffraction. In the present Section, we are concerned mostly with conical diffraction, including some notes about metamaterials. Classical diffraction is considered as a particular case with some important details for the implementation.

There exist different ways to transform the diffraction problems under consideration to one-dimensional integral equations over the boundary profile curve of the grating. It is beyond the scope of this chapter to describe the history of applying integral methods to grating problems and the variety of corresponding integral formulations. It should be mentioned that those methods were mostly developed by specialists in physics and optics, and, it seems, they were not aware of the rapid progress in the fields of "boundary integral equations" and "boundary element methods" made in the mathematical community since 1980.

12.2.1 Maxwell equations

We denote by $\mathbf{e}_x$, $\mathbf{e}_y$ and $\mathbf{e}_z$ the unit vectors of the axis of the Cartesian coordinates. The grating is a cylindrical surface whose generatrices are parallel to the z-axis (see Fig. 12.1) and whose cross section is described by the curve Σ (Fig. 12.2). We suppose that Σ is not self-intersecting and d-periodic in x-direction. The grating surface is the boundary between two regions $G_\pm \times \mathbb{R} \subset \mathbb{R}^3$ which are filled with materials of constant electric permittivity $\varepsilon_\pm$ and magnetic permeability $\mu_\pm$. We deal only with time-harmonic fields; consequently, the electric and magnetic fields are

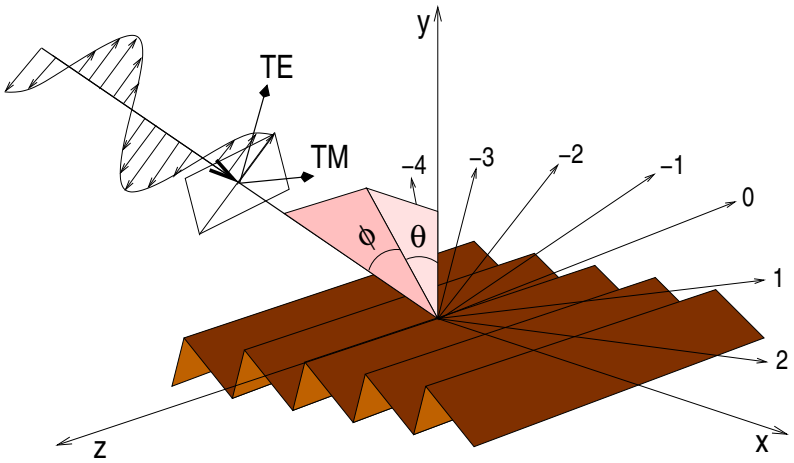


Figure 12.1: Schematic conical diffraction by a grating.

represented by the complex vectors $\mathbf{E}$ and $\mathbf{H}$, with a time dependence $\exp(-i\omega t)$ taken into account. The wave vector $\mathbf{k}_+$ of the incident wave in $G_+ \times \mathbb{R}$ is in general not perpendicular to the grooves ($\mathbf{k}_+ \cdot \mathbf{e}_z \neq 0$). Setting $\mathbf{k}_+ = (\alpha, -\beta, \gamma)$ the surface is illuminated by an electromagnetic plane wave

$$\mathbf{E}^i = \mathbf{p}\, \mathrm{e}^{i(\alpha x - \beta y + \gamma z)}\,, \quad \mathbf{H}^i = \mathbf{s}\, \mathrm{e}^{i(\alpha x - \beta y + \gamma z)}\,, \tag{12.1}$$

which due to the periodicity of Σ is scattered into a finite number of plane waves in $G_+ \times \mathbb{R}$ and possibly in $G_- \times \mathbb{R}$. The wave vectors of these outgoing modes lie on the surface of a cone whose axis is parallel to the z–axis. Therefore, one speaks of conical diffraction.

The components of $\mathbf{k}_+$ satisfy

$$\beta \in \mathbb{R} \quad \text{and} \quad \alpha^2 + \beta^2 + \gamma^2 = \omega^2 \varepsilon_+ \mu_+\,.$$

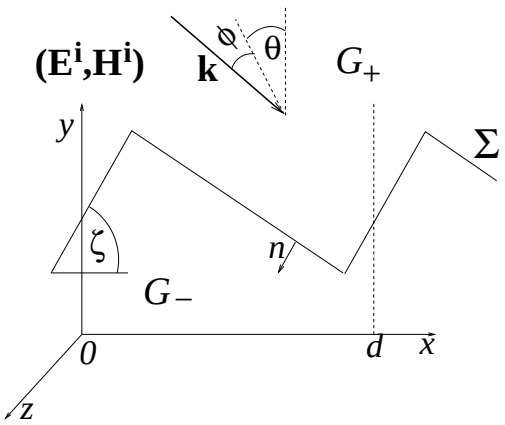


Figure 12.2: Cross section of a simple grating of period d with incidence direction $\mathbf{k}$, incidence angle θ and conical angle ϕ.

Note that this condition is satisfied by dielectric media with $\varepsilon_+ > 0, \mu_+ > 0$ as well as negative index materials, satisfying $\varepsilon_+ < 0, \mu_+ < 0$. The wave vector $\mathbf{k}_+$ is expressed using the incidence angles $|\theta|, |\phi| < \pi/2$

$$(\alpha, -\beta, \gamma) = \omega\sqrt{\varepsilon_+}\sqrt{\mu_+}\,(\sin\theta\cos\phi, -\cos\theta\cos\phi, \sin\phi)\,.$$

Note that $\beta > 0$ if $\varepsilon_+ > 0, \mu_+ > 0$, whereas $\beta < 0$ for negative index materials. In-plane diffraction corresponds to $\mathbf{k}_+ \cdot \mathbf{e}_z = 0$, the case $\phi \neq 0$ characterizes conical diffraction.

12.2.2 Helmholtz equations

Since the geometry is invariant with respect to any translation parallel to the z-axis, we make the ansatz for the total field

$$(\mathbf{E}, \mathbf{H})(x, y, z) = (E, H)(x, y)\, \mathrm{e}^{i\gamma z} \tag{12.2}$$

with the vector functions $E, H : \mathbb{R}^2 \to \mathbb{C}^3$. This transforms the time-harmonic Maxwell equations in $\mathbb{R}^3$

$$\boldsymbol{\nabla} \times \mathbf{E} = i\omega\mu\mathbf{H} \quad \text{and} \quad \boldsymbol{\nabla} \times \mathbf{H} = -i\omega\varepsilon\mathbf{E}\,, \tag{12.3}$$

with piecewise constant functions $\varepsilon(x, y) = \varepsilon_\pm$, $\mu(x, y) = \mu_\pm$ for $(x, y) \in G_\pm$, into a two-dimensional problem. Indeed, in regions with constant ε and μ we have the equations

$$\nabla_\gamma \times E = i\omega\mu H\,, \quad \nabla_\gamma \times H = -i\omega\varepsilon E\,, \quad \nabla_\gamma \cdot E = \nabla_\gamma \cdot H = 0\,, \tag{12.4}$$

with $\nabla_\gamma = (\partial_x, \partial_y, i\gamma)$. Then from (12.3)

$$\nabla_\gamma \times (\nabla_\gamma \times E) = \omega^2\varepsilon\mu E\,, \quad \nabla_\gamma \times (\nabla_\gamma \times H) = \omega^2\varepsilon\mu H\,. \tag{12.5}$$

Introducing the transverse components

$$E_T = E - E_z\mathbf{e}_z\,, \quad H_T = H - H_z\mathbf{e}_z\,,$$

we derive using (12.4)

$$\nabla_\gamma \times (\nabla_\gamma \times E_T) = \gamma^2 (E_T + E_z \mathbf{e}_z) + i\omega\mu \nabla \times (H_z \mathbf{e}_z)$$
$$\nabla_\gamma \times (\nabla_\gamma \times H_T) = \gamma^2 (H_T + H_z \mathbf{e}_z) - i\omega\varepsilon \nabla \times (E_z \mathbf{e}_z)$$

and

$$\nabla_\gamma \times (\nabla_\gamma \times E_z \mathbf{e}_z) = (i\gamma\partial_x E_z, i\gamma\partial_y E_z, -(\partial_x^2 + \partial_y^2) E_z)$$
$$\nabla_\gamma \times (\nabla_\gamma \times H_z \mathbf{e}_z) = (i\gamma\partial_x H_z, i\gamma\partial_y H_z, -(\partial_x^2 + \partial_y^2) H_z)$$

Thus, comparing the components in (12.4) we derive

$$\begin{aligned} (\omega^2\varepsilon\mu - \gamma^2) E_T &= i\gamma \nabla E_z + i\omega\mu \nabla \times (H_z \mathbf{e}_z)\,, \\ (\omega^2\varepsilon\mu - \gamma^2) H_T &= i\gamma \nabla H_z - i\omega\varepsilon \nabla \times (E_z \mathbf{e}_z)\,. \end{aligned} \tag{12.6}$$

We denote

$$\kappa^2 = \varepsilon\mu - \frac{\gamma^2}{\omega^2} = \varepsilon\mu - \varepsilon_+\mu_+ \sin^2\phi\,, \tag{12.7}$$

and conclude from (12.6) that under the condition $\kappa^2 \neq 0$, which will be assumed throughout, the components E_z, H_z determine the electromagnetic field $(\mathbf{E}, \mathbf{H})$.

Furthermore, comparing the third components we derive the relations

$$\omega^2\varepsilon\mu E_z = \gamma^2 E_z - (\partial_x^2 + \partial_y^2) E_z\,, \quad \omega^2\varepsilon\mu H_z = \gamma^2 H_z - (\partial_x^2 + \partial_y^2) H_z\,,$$

thus E_z and H_z are solutions of the two-dimensional Helmholtz equations in $G_\pm$

$$\Delta u + \omega^2\kappa^2 u = 0\,, \tag{12.8}$$

where $\Delta = \partial_x^2 + \partial_y^2$ denotes the Laplace operator in $\mathbb{R}^2$.

Denote by $\mathbf{n} = (n_x, n_y, 0)$ and $\mathbf{t} = \mathbf{e}_z \times \mathbf{n} = (-n_y, n_x, 0)$, respectively, the unit vectors of the normal and the tangent on the surface $\Gamma = \Sigma \times \mathbb{R}$. Then one finds from (12.6) that

$$\mathbf{n} \times E = \mathbf{n} \times E_T - E_z \mathbf{t} = \frac{i}{\omega^2\kappa^2} \Big(\gamma \mathbf{n} \times \nabla E_z + \omega\mu \mathbf{n} \times (\nabla \times (H_z \mathbf{e}_z)) \Big) - E_z \mathbf{t}$$

and

$$\mathbf{n} \times H = \mathbf{n} \times H_T - H_z \mathbf{t} = \frac{i}{\omega^2\kappa^2} \Big(\gamma \mathbf{n} \times \nabla H_z - \omega\varepsilon \mathbf{n} \times \nabla \times (E_z \mathbf{e}_z) \Big) - H_z \mathbf{t}$$

Thus the continuity of the tangential components $\mathbf{n} \times \mathbf{E}$ and $\mathbf{n} \times \mathbf{H}$ on the surface $\Gamma = \Sigma \times \mathbb{R}$ leads to the jump conditions for E_z, H_z across Σ of the form

$$\big[E_z\big]_\Sigma = \big[H_z\big]_\Sigma = 0\,, \ \Big[\frac{\gamma}{\omega^2\kappa^2}\,\partial_t H_z + \frac{\omega\varepsilon}{\omega^2\kappa^2}\,\partial_n E_z\Big]_\Sigma = \Big[\frac{\gamma}{\omega^2\kappa^2}\,\partial_t E_z - \frac{\omega\mu}{\omega^2\kappa^2}\,\partial_n H_z\Big]_\Sigma = 0\,. \tag{12.9}$$

Here $\partial_n = n_x\partial_x + n_y\partial_y$ and $\partial_t = -n_y\partial_x + n_x\partial_y$ are the normal and tangential derivatives on Σ, respectively, and $\big[u\big]_\Sigma$ denotes the jump of the function u across the curve Γ.

The z-components of the incoming field

$$E_z^i(x,y) = p_z\, \mathrm{e}^{i(\alpha x - \beta y)}\,, \ H_z^i(x,y) = s_z\, \mathrm{e}^{i(\alpha x - \beta y)} \tag{12.10}$$

are α-quasi-periodic in x of period d, i.e., they satisfy a Floquet condition

$$u(x+d,y) = \mathrm{e}^{id\alpha}\, u(x,y)\,.$$

In view of the periodicity of ε and μ this motivates to seek α-quasi-periodic solutions E_z, H_z.

Furthermore, the diffracted fields must remain bounded at infinity, which lead to the outgoing wave condition (OWC). If the profile curve Σ is contained in the strip $\{(x,y) : |y| < H\}$, then the quasiperiodicity of solutions implies outside the strip Rayleigh series expansion of the form

$$\begin{aligned}(E_z,H_z)(x,y) &= (E_z^i,H_z^i) + \sum_{n\in\mathbb{Z}} (E_n^+,H_n^+)\,\mathrm{e}^{i(\alpha_n x+\beta_n^+ y)}, \quad && y \geq H,\\ (E_z,H_z)(x,y) &= \sum_{n\in\mathbb{Z}} (E_n^-,H_n^-)\,\mathrm{e}^{i(\alpha_n x-\beta_n^- y)}, && y \leq -H,\end{aligned} \tag{12.11}$$

with unknown Rayleigh coefficients $E_n^\pm, H_n^\pm \in \mathbb{C}$, and α_n, $\beta_n^\pm$ given by the relations

$$\alpha_n = \alpha + \frac{2\pi n}{d}\,, \quad (\beta_n^\pm)^2 = \omega^2\kappa_\pm^2 - \alpha_n^2\,.$$

Then the functions are bounded for $|y| \to \infty$, if we choose the branch of the square root such that $\operatorname{Im}\beta_n^\pm \geq 0$, i. e. we set $z^{1/2} = r^{1/2}\exp(i\varphi/2)$ for $z = r\exp(i\varphi)$, $0 \leq \varphi < 2\pi$.

In the following it is always assumed that the material in G_+ satisfies either $\varepsilon_+, \mu_+ > 0$ or $\varepsilon_+, \mu_+ < 0$, and that the material parameters of the substrate are nonzero complex values with nonnegative imaginary part $\operatorname{Im}\varepsilon_-, \operatorname{Im}\mu_- \geq 0$. Thus, besides the usual optical materials also interesting negative index materials with $\varepsilon, \mu < 0$ are allowed.

If $0 \leq \arg\kappa_\pm^2 < 2\pi$, i.e. $\arg(\varepsilon_\pm + \mu_\pm) < 2\pi$, then $\beta_n^\pm$ with $0 \leq \arg\beta_n^\pm < \pi$ are properly defined for all n. However, if $\arg(\varepsilon_\pm + \mu_\pm) = 2\pi$, i.e. $\varepsilon_\pm, \mu_\pm < 0$, we set

$$\kappa_\pm = -\overline{\Big(\varepsilon_\pm\mu_\pm - \frac{\gamma^2}{\omega^2}\Big)^{1/2}}\,, \quad \beta_n^\pm = -\overline{\big(\omega^2\kappa_\pm^2 - \alpha_n^2\big)^{1/2}} \tag{12.12}$$

Summarizing, in case of off-plane diffraction the Maxwell system of equations reduces to two-dimensional Helmholtz equations for vector functions of two components (E_z, H_z), which satisfy the OWC (12.11) and are coupled by the jump conditions (12.9).

For in-plane diffraction ($\gamma = 0$) we derive from (12.9) the well-know fact, that one can consider the two fundamental cases of polarization separately, i.e. the TE mode (with the z component E_z of the electric field $\mathbf{E}$ parallel to the grating grooves) and the TM mode (with the z component H_z of the magnetic field $\mathbf{H}$ parallel to the grating grooves) (see Ch. 2).

Then the surface is illuminated by an electromagnetic plane wave

$$\mathbf{E}^i = \mathbf{p}\,\mathrm{e}^{i(\alpha x-\beta y)}\,, \quad \mathbf{H}^i = \mathbf{s}\,\mathrm{e}^{i(\alpha x-\beta y)}\,,$$

where $\alpha = k_+\sin\theta$, $\beta = k_+\cos\theta$, with the incidence angle $|\theta| < \pi/2$, and $k_+ = \omega^2\varepsilon_+\mu_+$ denotes the wavenumber inside $G_+ \times \mathbb{R}$.

The z-components of the total fields E_z (TE polarization) or H_z (TM polarization) satisfy the Helmholtz equation in $G_\pm$ except the boundary Σ

$$(\Delta + k_\pm^2)\,u_\pm = 0\,, \quad k_\pm^2 = \omega^2\varepsilon_\pm\mu_\pm \tag{12.13}$$

and satisfy the continuity conditions

$$u_+\big|_\Sigma = u_-\big|_\Sigma = 0,\ \partial_n u_+\big|_\Sigma = q\,\partial_n u_-\big|_\Sigma, \tag{12.14}$$

where $q = \mu_-/\mu_+$ or $q = \varepsilon_-/\varepsilon_+$ for the E_z- or H_z-component, respectively. In addition, they are subject to OWC (12.11) with $\beta_n^\pm$ given by relations

$$\beta_n^\pm = \sqrt{k_\pm^2 - \alpha_n^2}, \quad 0 \le \arg\beta_n^\pm < \pi. \tag{12.15}$$

Remark 12.2.1 *The results of this section can be easily generalized to more general diffraction gratings, where the outer domains G_+ and G_- are bounded by surfaces Σ_+ and Σ_-, surrounding an inner periodic grating structure. For that structure the d-periodic in x material parameters $\varepsilon(x,y)$ and $\mu(x,y)$ are piecewise continuous functions. Then the z-components E_z and H_z satisfy Helmholtz equations* (12.8) *with variable $\kappa^2 = \omega^2\varepsilon\mu$, where $\varepsilon(x,y)$ or $\mu(x,y)$ are continuous, and meet the jump conditions* (12.9) *at any discontinuous curve of $\varepsilon(x,y)$ or $\mu(x,y)$.*

12.2.3 Boundary integral operators

Here we describe the application of some mathematical results on boundary integral equations to the solution of the present Helmholtz problems in $G_\pm$. Let the common boundary Σ of G_- and G_+ be given by a piecewise C^2 parametrization

$$\sigma(t) = (X(t), Y(t)), \quad X(t+1) = X(t) + d,\ Y(t+1) = Y(t), \quad t \in \mathbb{R}, \tag{12.16}$$

i.e. the continuous functions X, Y are piecewise C^2 and $\sigma(t_1) \ne \sigma(t_2)$ if $t_1 \ne t_2$. If the profile Σ has corners, then we suppose additionally that the angles between adjacent tangents at the corners are strictly between 0 and 2π.

The potentials which provide α-quasi-periodic solutions of the Helmholtz equation

$$\Delta u + k^2 u = 0 \qquad \text{with } 0 \le \arg k^2 < 2\pi \tag{12.17}$$

are based on the quasi-periodic fundamental solution of period d

$$\Psi_{k,\alpha}(P) = \lim_{N\to\infty} \frac{i}{4} \sum_{n=-N}^{N} H_0^{(1)}\left(k\sqrt{(X-nd)^2+Y^2}\right) \mathrm{e}^{i\alpha nd}, \quad P = (X,Y), \tag{12.18}$$

with the Hankel function of the first kind $H_0^{(1)}$. The series (12.18) converges uniformly over compact sets in $\mathbb{R}^2 \setminus \bigcup_{n\in\mathbb{Z}} \{(nd, 0)\}$ if the condition

$$k^2 \ne \alpha_n^2 = \left(\alpha + \frac{2\pi n}{d}\right)^2 \quad \text{for all} \quad n \in \mathbb{Z} \tag{12.19}$$

is satisfied. At any point $Q = (nd, 0)$ the fundamental solution has a logarithmic singularity

$$\Psi_{k,\alpha}(P-Q) \asymp \frac{\mathrm{e}^{i\alpha(X-nd) - \sin(X-nd))}}{2\pi} \log\frac{1}{|P-Q|}$$

for $P = (X,Y)$ near Q. Moreover, setting $\beta_n = \sqrt{k^2 - \alpha_n^2}$ (recall that $\operatorname{Im}\beta_n \ge 0$) Poisson's summation formula leads to the representation

$$\Psi_{k,\alpha}(P) = \lim_{N\to\infty} \frac{i}{2d} \sum_{n=-N}^{N} \frac{\mathrm{e}^{i\alpha_n X + i\beta_n |Y|}}{\beta_n}, \tag{12.20}$$

where in the case $k<0$ according to (12.7) the fundamental solution is given by (12.20) with $\beta_n=-\overline{\left(k^2-\alpha_n^2\right)^{1/2}}$.

The single and double layer potentials are defined by

$$\mathscr{S}\varphi(P)=2\int_\Gamma \varphi(Q)\,\Psi_{k,\alpha}(P-Q)\,d\sigma_Q\,,\ \mathscr{D}\varphi(P)=2\int_\Gamma \varphi(Q)\,\partial_{n(Q)}\Psi_{k,\alpha}(P-Q)\,d\sigma_Q\,, \quad (12.21)$$

where Γ is one period of the interface Σ, i.e. $\Gamma=\{\sigma(t):t\in[t_0,t_0+1]\}$ for some t_0. In (12.21) $d\sigma_Q$ denotes the integration with respect to the arc length and $n(Q)$ is the normal to Σ at $Q\in\Sigma$ pointing into G_-. Obviously, for α-quasiperiodic densities φ on Σ the potentials $\mathscr{S}\varphi$, $\mathscr{D}\varphi$ are α-quasiperiodic in X and do not depend on the choice of Γ. They are solutions of the Helmholtz equation (12.17) in $G_\pm$ and satisfy the radiation condition

$$u(x,y)=\sum_{n=-\infty}^{\infty} u_n\,\mathrm{e}^{i\alpha_n x+i\beta_n|y|}\,,\quad |y|\geq H\,. \quad (12.22)$$

The potentials provide the usual representation formulas. Any α-quasiperiodic solution u of (12.17) in G_+ satisfying (12.22) admits the representation

$$\frac{1}{2}\left(\mathscr{S}\partial_n u-\mathscr{D}u\right)=\begin{cases} u & \text{in } G_+,\\ 0 & \text{in } G_-,\end{cases} \quad (12.23)$$

where the normal n points into G_-. Under the same assumptions for a function u in G_- the representation

$$\frac{1}{2}\left(\mathscr{D}u-\mathscr{S}\partial_n u\right)=\begin{cases} 0 & \text{in } G_+,\\ u & \text{in } G_-,\end{cases} \quad (12.24)$$

is valid.

Restriction of the potentials $\mathscr{S}$ and $\mathscr{D}$ to the profile curve Σ are the so called boundary integral operators. The potentials provide the usual jump relations of classical potential theory. The single layer potential is continuous across Σ

$$(\mathscr{S}\varphi)^+(P)=(\mathscr{S}\varphi)^-(P)=V\varphi(P)\,, \quad (12.25)$$

where the upper sign $+$ resp. $-$ denotes the limits of the potentials for points in $G_\pm$ tending in non-tangential direction to $P\in\Sigma$, and V is a integral operator with logarithmic singularity

$$V\varphi(P)=2\int_\Gamma \Psi_{k,\alpha}(P-Q)\,\varphi(Q)\,d\sigma_Q\,,\quad P\in\Sigma\,.$$

The double layer potential has a jump if crossing Γ:

$$\left(\mathscr{D}\varphi\right)^+=(K-I)\varphi,\quad \left(\mathscr{D}\varphi\right)^-=(K+I)\varphi \quad (12.26)$$

with the boundary double layer potential

$$K\varphi(P)=2\int_\Gamma \varphi(Q)\,\partial_{n(Q)}\Psi_{k,\alpha}(P-Q)\,d\sigma_Q+(\delta(P)-1)\varphi(P)\,.$$

Here $\delta(P)\in(0,2)$, $P\in\Sigma$, denotes the ratio of the angle in G_+ at P and π, i.e., $\delta(P)=1$ outside corner points of Σ. The normal derivative of $\mathscr{S}\varphi$ at Σ exists outside corners and has the limits

$$\left(\partial_n\mathscr{S}\varphi\right)^+=(L+I)\varphi,\quad \left(\partial_n\mathscr{S}\varphi\right)^-=(L-I)\varphi, \quad (12.27)$$

where L is the integral operator on Γ with the kernel $\partial_{n(P)}\Psi_{k,\alpha}(P-Q)$,

$$L\varphi(P) = 2\int_\Gamma \varphi(Q)\,\partial_{n(P)}\Psi_{k,\alpha}(P-Q)\,d\sigma_Q\,, \quad P\in\Sigma\,. \tag{12.28}$$

Boundary integral methods for second order partial differential equations usually employ also normal derivatives of the double layer potential. The kernel function of this integral operator has a strong singularity of the form $|P-Q|^{-2}$ near $Q=(nd,0)$, and must be interpreted as hyper-singular or integro-differential operator. Since the computation of this kernel function is rather complicated and time-consuming integral methods for diffraction gratings avoid this operator. However, in the following we need also the tangential derivative of single layer potentials

$$\partial_t\big(V\varphi\big)(P) = 2\,\partial_t\int_\Gamma \Psi_{k,\alpha}(P-Q)\,\varphi(Q)\,d\sigma_Q\,, \quad P\in\Sigma\,.$$

Interchanging differentiation and integration leads to an integral kernel with the non-integrable main singularity

$$\frac{t(P)\cdot(P-Q)}{|P-Q|^2}\,,$$

where $t(P)$ denotes the tangential vector to Σ at P. Therefore the tangential derivative of single layer potentials cannot be expressed as a usual integral. But it can be interpreted as the Cauchy principal value or singular integral

$$J\varphi(P) = 2\lim_{\delta\to0}\int_{\Gamma\setminus\Gamma(P,\delta)} \varphi(Q)\,\partial_{t(P)}\Psi_{k,\alpha}(P-Q)\,d\sigma_Q = \partial_t\big(V\varphi\big)(P)\,, \tag{12.29}$$

where $\Gamma(P,\delta)$ is the subarc of Γ of length 2δ with the mid point P. Similarly, one can define the singular integral

$$H\varphi(P) = 2\lim_{\delta\to0}\int_{\Gamma\setminus\Gamma(P,\delta)} \varphi(Q)\,\partial_{t(Q)}\Psi_{k,\alpha}(P-Q)\,d\sigma_Q\,, \tag{12.30}$$

which by using integration by parts gives for α-quasiperiodic φ

$$H\varphi(P) = -2\int_\Gamma \Psi_{k,\alpha}(P-Q)\,\partial_t\varphi(Q)\,d\sigma_Q = -V\big(\partial_t\varphi\big)(P)\,, \quad P\in\Sigma\,. \tag{12.31}$$

Note that $V\partial_t V = VJ = -HV$.

The integral formulation for the diffraction problems can be derived from the so-called direct or indirect approaches. The direct approach uses the representation formulas (12.23) or (12.24) together with the boundary values (12.25) and (12.26) to derive the boundary integral relations

$$V\partial_n u_+ - (I+K)u_+ = 0 \quad \text{or} \quad V\partial_n u_- + (I-K)u_- = 0 \tag{12.32}$$

for quasiperiodic solutions $u_\pm$ of the Helmholtz equations (12.17) in $G_\pm$ satisfying (12.22). Here the unknowns $u_\pm$ and $\partial_n u_\pm$ on the profile curve Γ have a direct physical meaning. For the indirect approach, the solution is sought as single or double layer potential with some unknown density.

An important ingredient for discussing the equivalence of integral formulations with the electromagnetic problem is the so-called "Uniqueness Theorem", which looks for conditions on Σ and the wave number k such that the solution of Helmholtz equation (12.17) in G_+ or G_- with zero boundary value on Σ is identical to zero in that domain. The uniqueness of this

Dirichlet problem is equivalent to the invertibility of the single layer potential operator V and is guaranteed by two sufficient conditions

– $\mathrm{Im}\, k^2 > 0$;

– the profile curve Σ satisfies $n_y(Q) \le 0$ for all $Q \in \Sigma$.

Hence, only gratings with overhanging profiles are not covered by the last condition, but it is a quite rare case that the Dirichlet problem with zero boundary data has a nontrivial solution. The only known examples, constructed in Ref. 12.15, are boundaries Σ of very exotic form, which will never appear in practice.

Therefore in the following we will always assume that the "Uniqueness Theorem" is valid. Then any quasiperiodic solution of the Helmholtz equation (12.17) in G_+ or G_- satisfying the OWC (12.22) can be uniquely determined via the representation formulas (12.23) or (12.24), respectively, or can be written as single layer potential $V\varphi$ with a quasiperiodic density φ, which belongs to some Sobolev-type space of functions given on Γ.

12.2.4 Integral equations for the in-plane case

Let us discuss examples of integral formulations for the in-plane diffraction case. Denote the z-components of the incident wave

$$u^i = \begin{cases} E_z^i & \text{for TE-polarization}, \\ H_z^i & \text{for TM-polarization}. \end{cases}$$

Then for bare (one-profile) gratings the problem (12.13) (12.14) (12.11) means that one has to find a solution of

$$\Delta u_\pm + k_\pm^2 u_\pm = 0 \quad \text{in } G_\pm, \tag{12.33}$$

satisfying continuity conditions on Σ

$$u_-|_\Sigma = (u_+ + u^i)|_\Sigma, \;\; \partial_n (u_+ + u^i)\big|_\Sigma = q\partial_n u_-\big|_\Sigma, \tag{12.34}$$

and the outgoing wave condition

$$\begin{aligned} u_+(x,y) &= \sum_{n=-\infty}^{\infty} c_n^+ \, \mathrm{e}^{i(\alpha_n x + \beta_n^+ y)} \qquad \text{for } y \ge H, \\ u_-(x,y) &= \sum_{n=-\infty}^{\infty} c_n^- \, \mathrm{e}^{i(\alpha_n x - \beta_n^- y)} \qquad \text{for } y \le -H. \end{aligned} \tag{12.35}$$

An example of the indirect approach is to look for densities φ_+ and φ_- on Γ such that

$$u_+ = \mathscr{S}^+ \varphi_+ \quad \text{and} \quad u_- = \mathscr{S}^- \varphi_- \tag{12.36}$$

satisfy (12.33-12.35), where $\mathscr{S}^\pm$ are the single layer potentials with the fundamental solution $\Psi_{k_\pm,\alpha}$. From (12.34),(12.25) and (12.27) one derives two integral equations on Γ

$$\begin{aligned} V^+\varphi_+ - V^-\varphi_- &= -u^i \\ (L^+ + I)\varphi_+ - q(L^- - I)\varphi_- &= -\partial_n u^i \end{aligned} \tag{12.37}$$

Here $L^\pm$ are defined by (12.28) with the kernel $\partial_{n(P)}\Psi_{k_\pm,\alpha}(P-Q)$.

The direct approach uses the relations

$$V^+\partial_n u_+ - (I+K^+)u_+ = 0\,,\; V^-\partial_n u_- + (I-K^-)u_- = 0\,, \tag{12.38}$$

following from (12.32), where the double layer potential $K^\pm$ has the kernel function $\partial_{n(Q)}\Psi_{k_\pm,\alpha}(P-Q)$. The integral equation in Ch. 4.2.3 was derived by putting (12.34) in the second equation in (12.38). Then one gets the system of integral equations

$$\begin{aligned} V^+\partial_n u_+ - (I+K^+)u_+ &= 0 \\ q^{-1}V^-\partial_n u_+ + (I-K^-)u_+ &= -q^{-1}V^-\partial_n u^i - (I-K^-)u^i \end{aligned} \tag{12.39}$$

for the unknowns u_+ and $\partial_n u_+$ as functions on Γ.

Another equation system with simpler right-hand side can be obtained by the direct method if one assumes that u^i is a solution of $\Delta u + k_+^2 u = 0$ in G_- and satisfies there (12.35). Hence, one gets additionally to (12.38)

$$V^+\partial_n u^i + (I-K^+)u^i = 0\,,$$

leading to the relation

$$V^+\partial_n(u_+ + u^i) - (I+K^+)(u_+ + u^i) = -2u^i\,. \tag{12.40}$$

As before, (12.34) implies two integral equations, but now for the unknowns u_- and $\partial_n u_-$

$$\begin{aligned} qV^+\partial_n u_- - (I+K^+)u_- &= -2u^i \\ V^-\partial_n u_- + (I-K^-)u_- &= 0 \end{aligned} \tag{12.41}$$

Note that both the direct and indirect approach lead for TE- and TM-polarization to a system of two linear integral equations with two unknowns. This is the standard approach in the boundary integral method for solving so-called transmission problems. Much effort has been spent in the theoretical and numerical analysis of different integral formulations and approximation methods for their effective solution.

It is popular in grating theory to combine the direct and indirect approaches, which results in a single integral equation for each polarization. This idea goes back to D. Maystre, who already in 1972 proposed this new approach (see Ch. 4). Take for example the representations

$$u_+ = \mathscr{S}^+\varphi_+ \text{ in } G_+ \quad \text{and} \quad u_- = \frac{1}{2}\big(\mathscr{D}^- u_- - \mathscr{S}^-\partial_n u_-\big) \text{ in } G_-\,.$$

Then by (12.34) and (12.27)

$$u_- = V^+\varphi_+ + u^i\,, \quad \partial_n u_- = q^{-1}\big((L^+ + I)\varphi_+ + \partial_n u^i\big)$$

such that the second relation in (12.38) implies the integral equation

$$\big(q^{-1}V^-(I+L^+) + (I-K^-)V^+\big)\varphi_+ = -\big(q^{-1}V^-\partial_n u^i + (I-K^-)u^i\big) \tag{12.42}$$

for one unknown density φ_+.

Another way is to set

$$u_+ = \frac{1}{2}\big(\mathscr{S}^+\partial_n u_+ - \mathscr{D}^+ u_+\big) \text{ in } G_+ \quad \text{and} \quad u_- = \mathscr{S}^-\varphi_- \text{ in } G_-\,.$$

In this case we derive from (12.34) and (12.27)

$$u_+ + u^i = V^- \varphi_- , \quad \partial_n (u_+ + u^i) = q(L^- - I)\varphi_- ,$$

such that (12.40) leads to the single integral equation

$$\big(qV^+(I-L^-) + (I+K^+)V^-\big)\varphi_- = 2u^i . \tag{12.43}$$

We see that the combined direct-indirect approach can lead to single integral equations for TE- and TM-problems on one-profile gratings. However, contrary to the pure direct or indirect integral method the equations contain products or compositions of boundary integral operators, which can lead to additional numerical difficulties.

12.2.5 Formulas for Rayleigh coefficients

After having solved one of the integral equation systems (12.37) or (12.41) or one of the single equations (12.42) or (12.43) it is easy to determine the complex amplitudes $c_n^\pm$ of the z-components of the plane waves (12.35) reflected and transmitted by the grating—the so-called Rayleigh coefficients. We note that the functions $u^\pm(X,Y)\,\mathrm{e}^{-i\alpha X}$ for fixed $\pm Y \geq H$, respectively, are smooth and d-periodic. Then $c_n^\pm e^{\pm i\beta_n^\pm Y}$ is simply the n-th Fourier coefficient of $u^\pm(X,Y)\,\mathrm{e}^{-i\alpha X}$, i.e.

$$c_n^\pm = \frac{\mathrm{e}^{\mp i\beta_n^\pm Y}}{d}\int_0^d u^\pm(X,Y)\,\mathrm{e}^{-i\alpha X}\,\mathrm{e}^{-2\pi nX/d}\,dX = \frac{\mathrm{e}^{\mp i\beta_n^\pm Y}}{d}\int_0^d u^\pm(X,Y)\,\mathrm{e}^{-i\alpha_n X}\,dX , \quad \pm Y \geq H .$$

The indirect approach leads to simple formulas. Suppose $u_\pm = \mathscr{S}^\pm\varphi_\pm$, with known density $\varphi_\pm$. Then for $(X,Y) = P$

$$\begin{aligned} c_n^\pm &= \frac{\mathrm{e}^{\mp i\beta_n^\pm Y}}{d}\int_0^d \mathscr{S}^\pm\varphi_\pm(P)\,\mathrm{e}^{-i\alpha_n X}\,dX \\ &= \frac{2\,\mathrm{e}^{\mp i\beta_n^\pm Y}}{d}\int_\Gamma \varphi_\pm(Q)\,d\sigma_Q \int_0^d \mathrm{e}^{-i\alpha_n X}\,\Psi_{k_\pm,\alpha}(P-Q)\,dX . \end{aligned}$$

It follows from (12.20) immediately, that with $Q = (x,y)$

$$\int_0^d \mathrm{e}^{-i\alpha_n X}\,\Psi_{k_\pm,\alpha}(P-Q)\,dX = \frac{i}{2}\frac{\mathrm{e}^{-i\alpha_n x + i\beta_n^\pm |Y-y|}}{\beta_n^\pm} \tag{12.44}$$

such that

$$c_n^\pm = \frac{i}{d\beta_n^\pm}\int_\Gamma \mathrm{e}^{-i\alpha_n x \mp i\beta_n^\pm y}\,\varphi_\pm(Q)\,d\sigma_Q , \quad \text{where} \quad Q = (x,y) . \tag{12.45}$$

Using the direct approach we find

$$u_\pm = \pm\frac{1}{2}\big(\mathscr{S}^\pm\varphi_\pm - \mathscr{D}^\pm\psi_\pm\big)$$

with known functions $\varphi_\pm, \psi_\pm$. Then

$$\begin{aligned} c_n^\pm &= \pm\frac{\mathrm{e}^{\mp i\beta_n^\pm Y}}{2d}\int_0^d \big(\mathscr{S}^\pm\varphi_\pm(P) - \mathscr{D}^\pm\psi_\pm(P)\big)\,\mathrm{e}^{-i\alpha_n X}\,dX \\ &= \frac{\mathrm{e}^{\mp i\beta_n^\pm Y}}{d}\int_\Gamma \big(\varphi_\pm(Q) - \psi_\pm(Q)\,\partial_{n(Q)}\big)\int_0^d \mathrm{e}^{-i\alpha_n X}\,\Psi_{k_\pm,\alpha}(P-Q)\,dX\,d\sigma_Q . \end{aligned}$$

Hence from (12.44) we get

$$c_n^{\pm} = \frac{i}{2d\beta_n^{\pm}} \int_\Gamma \left(\varphi_\pm(Q) - \psi_\pm(Q)\,\partial_n\right) \mathrm{e}^{-i\alpha_n x \mp i\beta_n^{\pm} y}\, d\sigma_Q\,. \tag{12.46}$$

Let us apply formulas (12.45), (12.46) to the single integral equation (12.43). The Rayleigh coefficients c_n^- can be determined from

$$c_n^- = \frac{i}{d\beta_n^-} \int_\Gamma \mathrm{e}^{-i\alpha_n x + i\beta_n^- y}\, \varphi_-(Q)\, d\sigma_Q\,, \tag{12.47}$$

where $Q = (x,y)$ and φ_- is the solution of (12.43). For the reflected waves we get

$$c_n^+ = \frac{i}{2d\beta_n^+} \int_\Gamma \Big(\big(q(L^- - I)\varphi_- - \partial_n u^i\big) - \big(V^-\varphi_- - u^i\big)\partial_n \Big) \mathrm{e}^{-i\alpha_n x - i\beta_n^+ y}\, d\sigma_Q\,. \tag{12.48}$$

12.2.6 Integral equations for the off-plane case

Now we describe the integral formulation of the conical diffraction (12.8), (12.49), (12.11) for one-profile gratings. To give the same physical dimension to the functions, we use the vacuum impedance $Z_v = (\mu_v/\varepsilon_v)^{1/2}$, where ε_v, μ_v denote the vacuum permittivity and permeability, respectively, and introduce $B_z = Z_v H_z$. Noting that $\gamma = \omega(\varepsilon_+\mu_+)^{1/2}\sin\phi$ the jump conditions (12.9) are rewritten in the form

$$\begin{gathered} \big[E_z\big]_\Sigma = \big[B_z\big]_\Sigma = 0\,, \\ \Big[\frac{\varepsilon\,\partial_n E_z}{\varepsilon_v\kappa^2}\Big]_\Sigma = -\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\,\Big[\frac{\partial_t B_z}{\kappa^2}\Big]_\Sigma, \quad \Big[\frac{\mu\,\partial_n B_z}{\mu_v\kappa^2}\Big]_\Sigma = \sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\,\Big[\frac{\partial_t E_z}{\kappa^2}\Big]_\Sigma. \end{gathered} \tag{12.49}$$

Denoting the z-components of the total fields

$$E_z = \begin{cases} u_+ + E_z^i \\ u_- \end{cases}, \quad B_z = \begin{cases} v_+ + B_z^i & \text{in } G_+, \\ v_- & \text{in } G_-\,, \end{cases}$$

the problem (12.8), (12.49), (12.11) can be written so as to find solutions of

$$\Delta u_\pm + \omega^2\kappa_\pm^2 u_\pm = \Delta v_\pm + \omega^2\kappa_\pm^2 v_\pm = 0 \quad \text{in } G_\pm, \quad \kappa_\pm^2 = \varepsilon_\pm\mu_\pm - \varepsilon_+\mu_+\sin^2\phi \tag{12.50}$$

having on Σ the jumps

$$\begin{aligned} &u_- = u_+ + E_z^i, \ \frac{\varepsilon_-\,\partial_n u_-}{\varepsilon_v\kappa_-^2} - \frac{\varepsilon_+\partial_n(u_+ + E_z^i)}{\varepsilon_v\kappa_+^2} = \sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\Big)\partial_t v_-, \\ &v_- = v_+ + B_z^i, \ \frac{\mu_-\,\partial_n v_-}{\mu_v\kappa_-^2} - \frac{\mu_+\partial_n(v_+ + B_z^i)}{\mu_v\kappa_+^2} = -\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\Big)\partial_t u_-, \end{aligned} \tag{12.51}$$

and satisfying the OWC

$$\begin{aligned} (u_+, v_+)(x,y) &= \sum_{n=-\infty}^{\infty} (E_n^+, B_n^+)\,\mathrm{e}^{i(\alpha_n x + \beta_n^+ y)} \qquad \text{for } y \ge H, \\ (u_-, v_-)(x,y) &= \sum_{n=-\infty}^{\infty} (E_n^-, B_n^-)\,\mathrm{e}^{i(\alpha_n x - \beta_n^- y)} \qquad \text{for } y \le -H. \end{aligned} \tag{12.52}$$

In order to represent $u_\pm$ and $v_\pm$ as layer potentials we assume in what follows that the parameters are such that $\beta_n^\pm = (\omega^2\kappa_\pm^2 - \alpha_n^2)^{1/2} \neq 0$ for all n. Similar to the combined approach of the previous Subsection 12.2.4 resulting in the very simple integral equation (12.43) the solutions in G_- are sought as single layer potentials

$$u_- = \mathscr{S}^- w\,, \quad v_- = \mathscr{S}^- \tau \tag{12.53}$$

with certain auxiliary densities w, τ, whereas the solutions u_+ and v_+ are expressed using (12.23), (12.24)

$$u_+ = \frac{1}{2}\big(\mathscr{S}^+\partial_n u_+ - \mathscr{D}^+ u_+\big), \qquad v_+ = \frac{1}{2}\big(\mathscr{S}^+\partial_n v_+ - \mathscr{D}^+ v_+\big) \qquad \text{in } G_+\,. \tag{12.54}$$

Here we denote by $\mathscr{S}^\pm$ the single layer potential defined on Γ with the fundamental solution $\Psi_{\omega\kappa_\pm,\alpha}$. Correspondingly $\mathscr{D}^\pm$ is the double layer potential over Γ with the normal derivative of $\Psi_{\omega\kappa_\pm,\alpha}$ as a kernel function. As in (12.40) we have

$$\begin{aligned} V^+\partial_n(u_+ + E_z^i) - \big(I + K^+\big)(u_+ + E_z^i) &= 2E_z^i|_\Sigma\,,\\ V^+\partial_n(v_+ + B_z^i) - \big(I + K^+\big)(v_+ + B_z^i) &= 2B_z^i|_\Sigma\,, \end{aligned} \tag{12.55}$$

where $V^\pm$ denote the boundary single layer potentials

$$V^\pm\varphi(P) = 2\int_\Gamma \varphi(Q)\,\Psi_{\omega\kappa_\pm,\alpha}(P-Q)\,d\sigma_Q\,, \quad P \in \Sigma\,,$$

and the operators $K^\pm$ and $L^\pm$ are defined analogously. Since by Eq. (12.27)

$$u_-|_\Sigma = V^- w,\ \partial_n u_-|_\Sigma = (L^- - I)w,\ v_-|_\Sigma = V^-\tau,\ \partial_n v_-|_\Sigma = (L^- - I)\tau,$$

we see from Eqs. (12.55) that the jump conditions (12.51) are valid when the unknowns w, τ satisfy the system of integral equations

$$\begin{aligned} &\frac{\varepsilon_-\kappa_+^2}{\varepsilon_+\kappa_-^2}V^+(L^- - I)w - \big(I + K^+\big)V^- w - \sqrt{\frac{\varepsilon_v\mu_+}{\varepsilon_+\mu_v}}\sin\phi\,\Big(1 - \frac{\kappa_+^2}{\kappa_-^2}\Big)V^+\partial_t V^-\tau = 2E_z^i,\\ &\frac{\mu_-\kappa_+^2}{\mu_+\kappa_-^2}V^+(L^- - I)\tau - \big(I + K^+\big)V^-\tau + \sqrt{\frac{\varepsilon_+\mu_v}{\varepsilon_v\mu_+}}\sin\phi\,\Big(1 - \frac{\kappa_+^2}{\kappa_-^2}\Big)V^+\partial_t V^- w = 2B_z^i. \end{aligned} \tag{12.56}$$

Recall that we suppose $\kappa_\pm^2 \neq 0$ and $\omega^2\kappa_\pm^2 - \alpha_n^2 \neq 0$ for all n.

For the analytical and numerical treatment of (12.56), it is advantageous to use the relations

$$V^+\partial_t V^- = -H^+V^- = V^+J^-$$

(see the definitions (12.29), (12.30)). Then (12.56) becomes a system of singular integral equations

$$\begin{aligned} &\frac{\varepsilon_-\kappa_+^2}{\varepsilon_+\kappa_-^2}V^+(I - L^-)w + (I + K^+)V^- w - \sqrt{\frac{\varepsilon_v\mu_+}{\varepsilon_+\mu_v}}\sin\phi\,\Big(1 - \frac{\kappa_+^2}{\kappa_-^2}\Big)H^+V^-\tau = -2E_z^i,\\ &\frac{\mu_-\kappa_+^2}{\mu_+\kappa_-^2}V^+(I - L^-)\tau + (I + K^+)V^-\tau + \sqrt{\frac{\varepsilon_+\mu_v}{\varepsilon_v\mu_+}}\sin\phi\,\Big(1 - \frac{\kappa_+^2}{\kappa_-^2}\Big)H^+V^- w = -2B_z^i. \end{aligned} \tag{12.57}$$

for which powerful analytical and numerical methods exist.

If the solution of the system (12.56) is found, then the solution of the conical diffraction problem (12.50)–(12.52) can be determined by the relations

$$u_+ = -\frac{1}{2}\Big(\mathscr{S}^+\Big(\frac{\varepsilon_-\kappa_+^2}{\varepsilon_+\kappa_-^2}(I-L^-)w + \sqrt{\frac{\varepsilon_v\mu_+}{\varepsilon_+\mu_v}}\sin\phi\,\big(1-\frac{\kappa_+^2}{\kappa_-^2}\big)J^-\tau + \partial_n E_z^i\Big) + \mathscr{D}^+V^-(w-E_z^i)\Big),$$

$$v_+ = -\frac{1}{2}\Big(\mathscr{S}^+\Big(\frac{\mu_-\kappa_+^2}{\mu_+\kappa_-^2}(I-L^-)\tau - \sqrt{\frac{\varepsilon_+\mu_v}{\varepsilon_v\mu_+}}\sin\phi\,\big(1-\frac{\kappa_+^2}{\kappa_-^2}\big)J^-w + \partial_n B_z^i\Big) + \mathscr{D}^+V^-(\tau-B_z^i)\Big),$$

$$u_- = \mathscr{S}^-w, \quad v_- = \mathscr{S}^-\tau. \tag{12.58}$$

A detailed mathematical analysis of the system (12.56) is given in Ref. 12.16. In particular, the following properties have been established:

1. The integral equations are equivalent to the Helmholtz system if the operators V^+ and V^- are invertible.
2. If the profile Σ has no corners, then (12.56) is solvable if $\varepsilon_- + \varepsilon_+ \neq 0$ and $\mu_- + \mu_+ \neq 0$.
3. If the profile Σ has corners, then (12.56) is solvable if $\varepsilon_-/\varepsilon_+$ and $\mu_-/\mu_+ \notin [-\rho, -1/\rho]$ for some $\rho > 1$, depending on the angles at these corners.
4. The solution of (12.56) is unique if $\operatorname{Im}\varepsilon_- \geq 0$ and $\operatorname{Im}\mu_- \geq 0$ with $\operatorname{Im}(\varepsilon_- + \mu_-) > 0$

Remark 12.2.2 *The one-boundary solver can be used effectively in multilayer grating problems with separating boundaries, i.e., the maximal y value of a given profile is strictly less than the minimal y value of the next profile above (see Se. 12.5.1). In this case, it is possible to determine the diffracted field of the grating by computing scattering amplitude matrices separately for any profile. For each interface between two different materials, the computation of the scattering amplitude matrices corresponds to solving one-boundary conical diffraction problems with plane waves illuminating the interface from above and below.*

12.3 Efficiency, absorption, and energy balance

In this part, we give formulas for efficiencies and the absorption of bare gratings under oblique incidences and discuss the energy balance.

12.3.1 Efficiencies in conical diffraction

Diffraction efficiencies or far field patterns for the reflected and transmitted fields can easily be found from the corresponding Rayleigh coefficients of the diffracted outgoing waves. Defining the "energy" as the flux of Poynting's vector

$$\mathbf{P} = \operatorname{Re}(\mathbf{E} \times \overline{\mathbf{H}})/2 \tag{12.59}$$

through a normed rectangle parallel to the (x,z)-plane, the ratio of the energies of a reflected or transmitted propagating mode and of the incident wave is defined as the efficiency of that diffracted order. Thus, for a propagating plane wave $(\mathbf{E},\mathbf{H}) = (\mathbf{p},\mathbf{s})\,\mathrm{e}^{i(k_x x + k_y y + k_z z)}$, $\mathbf{k} = (k_x, k_y, k_z)$ with $|\mathbf{k}|^2 = \omega^2\varepsilon\mu$, the energy is proportional to the y-component of Poynting's vector

$$\mathbf{P}_y = \frac{1}{2}\operatorname{Re}(p_z\overline{s_x} - p_x\overline{s_z}).$$

Since by (12.6)

$$p_x = -\frac{1}{|\mathbf{k}|^2 - k_z^2}\left(k_x k_z p_z + \omega\mu k_y s_z\right),\ s_x = \frac{1}{|\mathbf{k}|^2 - k_z^2}\left(\omega\varepsilon k_y p_z - k_x k_z s_z\right)$$

we obtain

$$\mathbf{P}_y = \frac{\omega k_y}{2(|\mathbf{k}|^2 - k_z^2)}(\varepsilon|p_z|^2 + \mu|s_z|^2) = \frac{\omega\varepsilon_v k_y}{2(|\mathbf{k}|^2 - k_z^2)}\Big(\frac{\varepsilon}{\varepsilon_v}|p_z|^2 + \frac{\mu}{\mu_v}|q_z|^2\Big) \tag{12.60}$$

for $(\mathbf{E},\mathbf{B}) = (\mathbf{p},\mathbf{q})\,\mathrm{e}^{i(k_x x + k_y y + k_z z)}$ with the modification $\mathbf{B} = (\mu_v/\varepsilon_v)^{1/2}\mathbf{H}$.

To find relations between the efficiencies in conical diffraction, let E_z, B_z be a solution of the partial differential formulation of conical diffraction (12.8), (12.49) and (12.11). The expression of the conservation of energy can be derived from a variational equality for E_z and B_z in a periodic cell Ω_H, which has in x-direction the width d, is bounded by the straight lines $\{y = \pm H\}$ and contains Γ. We multiply Eqs.

$$(\Delta + \omega^2\kappa^2)\,E_z = (\Delta + \omega^2\kappa^2)\,B_z = 0$$

in $G_\pm$. respectively with

$$\frac{\varepsilon}{\varepsilon_v\kappa^2}\,\overline{E_z} \quad \text{and} \quad \frac{\mu}{\mu_v\kappa^2}\,\overline{B_z},$$

and apply Green's formula in the subdomains $\Omega_H \cap G_\pm$. Then by using the quasi-periodicity of E_z, B_z and the jump relations (12.49), one can derive

$$\begin{aligned}
\int_{\Omega_H}\frac{\varepsilon}{\varepsilon_v}\Big(\frac{1}{\kappa^2}|\nabla E_z|^2 - \omega^2|E_z|^2\Big) &+ \sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\Big(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\Big)\int_\Gamma \partial_t B_z\,\overline{E_z} \\
&- \frac{\varepsilon_+}{\varepsilon_v\kappa_+^2}\int_{\Gamma(H)}\partial_n E_z\,\overline{E_z} - \frac{\varepsilon_-}{\varepsilon_v\kappa_-^2}\int_{\Gamma(-H)}\partial_n E_z\,\overline{E_z} = 0\,, \qquad (12.61)\\
\int_{\Omega_H}\frac{\mu}{\mu_v}\Big(\frac{1}{\kappa^2}|\nabla B_z|^2 - \omega^2|B_z|^2\Big) &- \sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\Big(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\Big)\int_\Gamma \partial_t E_z\,\overline{B_z} \\
&- \frac{\mu_+}{\mu_v\kappa_+^2}\int_{\Gamma(H)}\partial_n B_z\,\overline{B_z} - \frac{\mu_-}{\mu_v\kappa_-^2}\int_{\Gamma(-H)}\partial_n B_z\,\overline{B_z} = 0\,, \qquad (12.62)
\end{aligned}$$

where $\Gamma(\pm H)$ denotes the upper and lower straight boundary of Ω_H, respectively, and the normal n on $\Gamma(\pm H)$ is directed outward. The outgoing wave condition (12.11) implies

$$\begin{aligned}
\int_{\Gamma(H)}\partial_n E_z\,\overline{E_z} &= i\beta\Big(|E_0^+|^2 - |p_z|^2 + 2i\,\mathrm{Im}\left(E_0^+\,\overline{p_z}\,\mathrm{e}^{i\beta H}\right)\Big) + i\sum_{n\neq 0}\beta_n^+|E_n^+|^2\,\mathrm{e}^{-2H\,\mathrm{Im}\,\beta_n^+}\,,\\
\int_{\Gamma(-H)}\partial_n E_z\,\overline{E_z} &= i\sum_{n\in\mathbb{Z}}\beta_n^-|E_n^-|^2\,\mathrm{e}^{-2H\,\mathrm{Im}\,\beta_n^-}\,,
\end{aligned} \tag{12.63}$$

and similar expressions for the boundary integrals involving B_z.

Note that ε_+ and μ_+ are nonzero real numbers, and let ε_- and μ_- also be real. Taking the

imaginary part of Eqs. (12.61) and (12.62) one gets

$$\begin{aligned}
&\frac{\varepsilon_+}{\varepsilon_v\kappa_+^2}\beta|p_z|^2-\frac{\varepsilon_+}{\varepsilon_v\kappa_+^2}\sum_{\beta_n^+>0}\beta_n^+|E_n^+|^2-\frac{\varepsilon_-}{\varepsilon_v\kappa_-^2}\sum_{\beta_n^->0}\beta_n^-|E_n^-|^2\\
&\qquad=-\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\Big(\frac{1}{\kappa_+^2}-\frac{1}{\kappa_-^2}\Big)\operatorname{Im}\int_\Gamma\partial_t B_z\overline{E_z}\,,\\
&\frac{\mu_+}{\mu_v\kappa_+^2}\beta|q_z|^2-\frac{\mu_+}{\mu_v\kappa_+^2}\sum_{\beta_n^+>0}\beta_n^+|B_n^+|^2-\frac{\mu_-}{\mu_v\kappa_-^2}\sum_{\beta_n^->0}\beta_n^-|B_n^-|^2\\
&\qquad=\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\Big(\frac{1}{\kappa_+^2}-\frac{1}{\kappa_-^2}\Big)\operatorname{Im}\int_\Gamma\partial_t E_z\overline{B_z}\,,
\end{aligned}$$

which in view of

$$\operatorname{Im}\int_\Gamma\partial_t B_z\overline{E_z}=\operatorname{Im}\int_\Gamma\partial_t E_z\overline{B_z}$$

leads to

$$\begin{aligned}
&\frac{\beta}{\kappa_+^2}\Big(\frac{\varepsilon_+}{\varepsilon_v}|p_z|^2+\frac{\mu_+}{\mu_v}|q_z|^2\Big)\\
&=\sum_{\beta_n^+>0}\frac{\beta_n^+}{\kappa_+^2}\Big(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\Big)+\sum_{\beta_n^->0}\frac{\beta_n^-}{\kappa_-^2}\Big(\frac{\varepsilon_-}{\varepsilon_v}|E_n^-|^2+\frac{\mu_-}{\mu_v}|B_n^-|^2\Big)\,.
\end{aligned}\tag{12.64}$$

Comparing with (12.60), we see that (12.64) relates the energy of the incident wave, which is proportional to the left side of (12.64) with the energies of the reflected and transmitted modes

$$\frac{\beta_n^+}{\kappa_+^2}\Big(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\Big)\quad\text{and}\quad\frac{\beta_n^-}{\kappa_-^2}\Big(\frac{\varepsilon_-}{\varepsilon_v}|E_n^-|^2+\frac{\mu_-}{\mu_v}|B_n^-|^2\Big),\tag{12.65}$$

respectively. Thus, setting the energy of the incident wave

$$\frac{\varepsilon_+}{\varepsilon_v}|p_z|^2+\frac{\mu_+}{\mu_v}|q_z|^2=1\,,\tag{12.66}$$

from (12.64) we derive for lossless gratings that $R+T=1$, where R denotes the sum of reflection order efficiencies

$$R=\sum_{\beta_n^+>0}\frac{\beta_n^+}{\beta}\Big(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\Big)=\sum_{\beta_n^+>0}\eta_n^+\tag{12.67}$$

and T is the sum of transmission order efficiencies

$$T=\frac{\kappa_+^2}{\kappa_-^2}\sum_{\beta_n^->0}\frac{\beta_n^-}{\beta}\Big(\frac{\varepsilon_-}{\varepsilon_v}|E_n^-|^2+\frac{\mu_-}{\mu_v}|B_n^-|^2\Big)=\sum_{\beta_n^+>0}\eta_n^-\,.\tag{12.68}$$

Here, the Rayleigh coefficients $E_n^\pm$ and $B_n^\pm$ can be derived using the formulas presented in Section 12.2.5. For example, (12.45) and (12.58) in G_- lead to

$$E_n^-=\frac{i}{d\beta_n^-}\int_\Gamma \mathrm{e}^{-i\alpha_n x+i\beta_n^- y}\,w(Q)\,d\sigma_Q\,,\quad B_n^-=\frac{i}{d\beta_n^-}\int_\Gamma \mathrm{e}^{-i\alpha_n x+i\beta_n^- y}\,\tau(Q)\,d\sigma_Q\,,\tag{12.69}$$

with $(x,y) = Q \in \Gamma$. Further, the direct integral representation in G_+ implies

$$E_+ = \frac{1}{2}\big(\mathscr{S}^+\varphi_E - \mathscr{D}^+\psi_E\big)\,, \quad B_+ = \frac{1}{2}\big(\mathscr{S}^+\varphi_B - \mathscr{D}^+\psi_B\big)\,,$$

with the known functions (cf. (12.58))

$$\varphi_E = -\left(\frac{\varepsilon_-\kappa_+^2}{\varepsilon_+\kappa_-^2}(I-L^-)w + \sqrt{\frac{\varepsilon_v\mu_+}{\varepsilon_+\mu_v}}\sin\phi\left(1-\frac{\kappa_+^2}{\kappa_-^2}\right)J^-\tau + \partial_n E_z^i\right), \quad \psi_E = V^-(w-E_z^i)\,.$$
$$\varphi_B = -\left(\frac{\mu_-\kappa_+^2}{\mu_+\kappa_-^2}(I-L^-)\tau - \sqrt{\frac{\varepsilon_+\mu_v}{\varepsilon_v\mu_+}}\sin\phi\left(1-\frac{\kappa_+^2}{\kappa_-^2}\right)J^-w + \partial_n B_z^i\right), \quad \psi_B = V^-(\tau-B_z^i)\,.$$

Thus, from (12.46)

$$\begin{aligned} E_n^+ &= \frac{i}{2d\beta_n^+}\int_\Gamma \big(\varphi_E(Q) - \psi_E(Q)\,\partial_n\big)\,\mathrm{e}^{-i\alpha_n x \mp i\beta_n^\pm y}\,d\sigma_Q\,, \\ B_n^+ &= \frac{i}{2d\beta_n^+}\int_\Gamma \big(\varphi_B(Q) - \psi_B(Q)\,\partial_n\big)\,\mathrm{e}^{-i\alpha_n x \mp i\beta_n^\pm y}\,d\sigma_Q\,. \end{aligned} \tag{12.70}$$

12.3.2 Generalization of energy balance for absorbing bare gratings

One of the most important accuracy criteria based on a single computation is the energy balance that can be generalized in the lossy bulk case described in this Subsection. If the grating is perfectly conducting, then the conservation of energy is expressed by the standard energy criterion

$$R = 1,$$

where R is the sum of the reflection order efficiencies.

If the grating is lossless, $\operatorname{Im}\varepsilon_- = 0$ and $\operatorname{Im}\mu_- = 0$, then conservation of energy is expressed by a similar energy criterion (see (12.67) and (12.68))

$$R + T = 1,$$

where T is the sum of the transmission order efficiencies.

In a general case, if $\operatorname{Im}\varepsilon_- \neq 0$ or $\operatorname{Im}\mu_- \neq 0$, then $T = 0$, $R < 1$, and the remaining part A of the energy is absorbed in the substrate

$$A + R = 1. \tag{12.71}$$

A is called the absorption coefficient or simply the absorption in the given diffraction problem. Therefore, an important tool to check the quality of the numerical solution for absorbing gratings is the requirement that the sum of the reflected energy and the absorption energy should be equal to the energy of the incident wave. Besides being physically meaningful, the expression (12.71) is very useful as one of numerical accuracy tests for computational codes and especially important in the x-ray and EUV ranges, and also for plasmonics and metamaterials applications, where absorption plays a predominant role. In the lossy case, one needs an independently calculated quantity A to verify (12.71). For such a quantity, we use the absorption integral defined in Ref. 12.7 and derived bellow.

12.3.3 Absorption for bare gratings

To obtain an expression for the absorption energy we apply Green's formula to E_z and B_z in the domain $\Omega_H \cap G_+$, which gives, since the normal n on Γ is exterior for $\Omega_H \cap G_+$

$$\begin{aligned}\int_{\Omega_H\cap G_+}\left(|\nabla E_z|^2-\omega^2\kappa_+^2\,|E_z|^2\right)&=\int_{\Gamma(H)}\partial_n E_z\,\overline{E_z}+\int_{\Gamma}\partial_n E_z\,\overline{E_z}\,,\\ \int_{\Omega_H\cap G_+}\left(|\nabla B_z|^2-\omega^2\kappa_+^2\,|B_z|^2\right)&=\int_{\Gamma(H)}\partial_n E_z\,\overline{E_z}+\int_{\Gamma}\partial_n E_z\,\overline{E_z}\,.\end{aligned}\tag{12.72}$$

The outgoing wave condition (12.11) imply (12.63), such that

$$\mathrm{Im}\int_{\Gamma(H)}\partial_n E_z\,\overline{E_z}=-\beta\,|p_z|^2+\sum_{\beta_n^+\geq 0}\beta_n^+|E_n^+|^2\,,\quad \mathrm{Im}\int_{\Gamma(H)}\partial_n B_z\,\overline{B_z}=-\beta\,|q_z|^2+\sum_{\beta_n^+\geq 0}\beta_n^+|B_n^+|^2$$

Since κ_+ is real, the imaginary parts on the left of Eqs. 12.72 vanish, and therefore

$$\frac{\varepsilon_+}{\varepsilon_v}|p_z|^2+\frac{\mu_+}{\mu_v}|q_z|^2=\sum_{\beta_n^+\geq 0}\frac{\beta_n^+}{\beta}\left(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\right)+\frac{\varepsilon_+}{\varepsilon_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n E_z\,\overline{E_z}+\frac{\mu_+}{\mu_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n B_z\,\overline{B_z}\,.$$

Thus, setting as before in (12.66) the energy of the incident wave

$$\frac{\varepsilon_0}{\varepsilon_v}|p_z|^2+\frac{\mu_0}{\mu_v}|q_z|^2=1\,,$$

the sum of reflection order efficiencies R fulfils

$$R+\frac{\varepsilon_+}{\varepsilon_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n E_z\,\overline{E_z}+\frac{\mu_+}{\mu_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n B_z\,\overline{B_z}=1\,,$$

i.e., we derive the conservation of energy for absorbing gratings $R+A=1$ with the absorption

$$A=\frac{\varepsilon_+}{\varepsilon_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n E_z\,\overline{E_z}+\frac{\mu_+}{\mu_v\beta}\,\mathrm{Im}\int_\Gamma\partial_n B_z\,\overline{B_z}\,.\tag{12.73}$$

Note that $\partial_n E_z=\partial_n^+ E_z$ and $\partial_n B_z=\partial_n^+ B_z$ are the normal derivatives on Γ of the z-components of the total fields in G_+, i.e. the sum of the reflected and the incident fields. Using the jump condition (12.51) the formula for A in the case of conical diffraction can be written in the form

$$A=\frac{\kappa_+^2}{\beta}\,\mathrm{Im}\left(\frac{1}{\kappa_-^2}\left(\frac{\varepsilon_-}{\varepsilon_v}\int_\Gamma\partial_n^- E_z\,\overline{E_z}+\frac{\mu_-}{\mu_v}\int_\Gamma\partial_n^- B_z\,\overline{B_z}+2\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\sin\phi\,\mathrm{Re}\int_\Gamma E_z\,\partial_t\overline{B_z}\right)\right).$$

In terms of the solution w,τ of the integral equations (12.56) the absorption energy is given by the formula

$$\begin{aligned}A=\frac{\kappa_+^2}{\beta}\,\mathrm{Im}\left(\frac{1}{\kappa_-^2}\int_\Gamma\left(\frac{\varepsilon_-}{\varepsilon_v}(L^--I)w\,\overline{V^-w}+\frac{\mu_-}{\mu_v}(L^--I)\tau\,\overline{V^-\tau}\right)\right)\\ +\frac{2\kappa_+^2\sin\phi}{\beta}\sqrt{\frac{\varepsilon_+\mu_+}{\varepsilon_v\mu_v}}\,\mathrm{Im}\,\frac{1}{\kappa_-^2}\,\mathrm{Re}\int_\Gamma V^-w\,\overline{J^-\tau}\,.\end{aligned}\tag{12.74}$$

12.3.4 Efficiencies and absorption for in-plane diffraction

In the special case of in-plane diffraction ($\phi = 0$) these formulas provide for lossless gratings

$$\begin{aligned}&\frac{\varepsilon_+}{\varepsilon_v}|p_z|^2+\frac{\mu_+}{\mu_v}|q_z|^2\\&=\sum_{\beta_n^+>0}\frac{\beta_n^+}{\beta}\Big(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\Big)+\sum_{\beta_n^-<0}\frac{\beta_n^-}{\beta}\Big(\frac{\mu_+}{\mu_-}\frac{\varepsilon_+}{\varepsilon_v}|E_n^-|^2+\frac{\varepsilon_+}{\varepsilon_-}\frac{\mu_+}{\mu_v}|B_n^-|^2\Big),\end{aligned}\tag{12.75}$$

and for absorbing gratings we derive the relation

$$\begin{aligned}&\frac{\varepsilon_+}{\varepsilon_v}|p_z|^2+\frac{\mu_+}{\mu_v}|q_z|^2\\&=\sum_{\beta_n^+>0}\frac{\beta_n^+}{\beta}\Big(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2+\frac{\mu_+}{\mu_v}|B_n^+|^2\Big)+\frac{1}{\beta}\,\mathrm{Im}\left(\frac{\mu_+}{\mu_-}\frac{\varepsilon_+}{\varepsilon_v}\int_\Gamma\partial_n E_z\overline{E_z}+\frac{\varepsilon_+}{\varepsilon_-}\frac{\mu_+}{\mu_v}\int_\Gamma\partial_n B_z\overline{B_z}\right).\end{aligned}\tag{12.76}$$

We get the well-known expressions for efficiencies $\eta_n^\pm$ and the heat absorption for TE and TM polarizations

$$\eta_n^+(TE)=\frac{\beta_n^+}{\beta}\frac{|c_n^+(TE)|^2}{|E_z^i|^2},\qquad \eta_n^+(TM)=\frac{\beta_n^+}{\beta}\frac{|c_n^+(TM)|^2}{|H_z^i|^2},$$

$$\eta_n^-(TE)=\frac{\mu_+}{\mu_-}\frac{\beta_n^-}{\beta}\frac{|c_n^-(TE)|^2}{|E_z^i|^2},\qquad \eta_n^-(TM)=\frac{\varepsilon_+}{\varepsilon_-}\frac{\beta_n^+}{\beta}\frac{|c_n^-(TM)|^2}{|H_z^i|^2},$$

$$A(TE)=\frac{1}{\beta}\,\mathrm{Im}\,\frac{\mu_+}{\mu_-}\int_\Gamma\partial_n E_z\overline{E_z},\qquad A(TM)=\frac{1}{\beta}\,\mathrm{Im}\,\frac{\varepsilon_+}{\varepsilon_-}\int_\Gamma\partial_n H_z\overline{H_z}.$$

12.4 Numerical solution of single-boundary problems

Progress in algorithms for numerical solutions of 2D and 3D Helmholtz equations in the last two decades has been nearly comparable with that of computer hardware and experimental nanophotonics. Note that some numerical methods, e.g. differential and CWA, but not integral, inherently suffer from ill-conditionness as wavenumbers or order numbers increase. However, algorithms that possess superior convergence properties are less universal with respect to a scatterer geometry and not well-behaved in the high-frequency range. By these reasons, in the present commercial and non-commercial codes based on the IMs, more classical and robust approaches are mostly used, however with a few possible modifications proposed for low λ/d ratio problems and described in Sec. 12.7.

In practice, the convergence and accuracy of efficiency computation using IMs depend significantly on a proper choice of discretization schemes, quadrature rules, and summation methods for the computation of integral kernels. In order to additionally reduce time (up to an order for some problems) for computation matrices of the above operator equations algorithmic enhancements can be applied to one-boundary solvers. It can be done by using, e.g., cache for exponential functions (plane waves) and cache for kernel functions, as described in the following.

12.4.1 Mathematical results for the integral equations

Here we point out some important mathematical aspects of the integral equations (12.42), (12.43) or (12.57), which are the basis for the described IMs. From the theoretical point of

view these equations are almost invertible as operators acting between different Sobolev spaces. If the right-hand side of the integral equation has certain smoothness, say r, i.e. it belongs to a Sobolev space H^r, then the solution has smoothness $r-1$ and belongs to H^{r-1}. The correct choice of the Sobolev spaces and the parameter r is well understood; it depends on the smoothness of the boundary profile.

Informally speaking, the equations correspond to linear operators of the form

$$(aI+bK)V+C$$

where V and K are single and double layer potentials with an arbitrary wavenumber including $k=0$, a and b are some constants, and C is compact in the pairs of Sobolev spaces $H^{r-1}\to H^r$. The single layer potential V is always invertible as operator $H^{r-1}\to H^r$. If the profile is smooth, then K is compact in $H^r\to H^r$, hence the equations generate an operator

$$aV+C_1\,,$$

which is invertible for almost all parameters in the pairs of Sobolev spaces for all r. Then any results on single layer potentials equations and their approximate solution can be applied.

If the profile has corners, then V is bounded and invertible as operator $H^{r-1}\to H^r$ for $0\le r\le 1$, K is not compact in $H^r\to H^r$, but bounded in this range; hence one has to study the above operator in the pairs of Sobolev spaces $H^{r-1}\to H^r$ for $0\le r\le 1$. It is invertible except a discrete number of parameter values and one must apply discretization results for both V and for $aI+bK$. One serious problem arises that the solution is in general not in L_2, it has singularities of the form $O(\rho^{-\delta})$, $0<\delta<1$, where ρ is the distance to the closest edge. Then the solution is not finite at the corner points.

This and some other pure mathematical problems will not be discussed in this Chapter, however they have great influence in practice and should be mentioned here. Concerning the numerical solution of the integral equations (12.42), (12.43) or (12.57), one has to consider at least three important theoretical problems:

1. The discretization of V,K,L,H and their products;
2. The convergence of the chosen numerical method under the condition that the kernel functions are exact or computed with some tolerance;
3. The efficient computation of the kernel functions with given accuracy.

For many types of integral operators quite satisfactory solutions to these problems are known, but not all of them have been applied to the diffraction integrals under consideration.

12.4.2 Approximation of integral equations

Here we describe briefly special Nyström and collocation methods used by the authors for solving the equations (12.42), (12.43) or (12.57). In the described realizations, they are rather simple but robust and universal methods with possibly a small number of kernel computations, because this procedure is rather expensive for the diffraction integrals. The discretization of the products of the integral operators appearing in these equations is done with the separate discretization of the integrals

$$Ax(t)=\upsilon x(t)+\int_0^1 K(t,s)x(s)ds\,. \tag{12.77}$$

Here we use the parametrization (12.16). Neglecting the factor $e^{i\alpha X(t)}$ the kernel $K(t,s)$ and the unknown function $x(s)$ are periodic. The case $\upsilon = \pm 1$ corresponds to the integrals with K and L, whereas $\upsilon = 0$ for the integral operators with singular kernels V, H and J.

The Nyström discretization of A is based on a quadrature rule of the integral

$$\int_0^1 x(s)ds \approx \sum_{j=1}^{N} w_j x(t_j)\,,$$

where for periodic functions already the rectangular rule $w_j = q$, $t_j = jq$ with $q = 1/N$ provides exponential convergence for smooth functions x. The solution $\{\sigma_k\}_{k=1}^N$ of the linear system

$$\upsilon\sigma_k + q\sum_{j=1}^{N} K(t_k,t_j)\sigma_j = y(t_k)\,, \qquad \text{for all } t_k\,,\ k = 1,\ldots,N\,,$$

is the Nyström approximate solution of the equation $Ax(t) = y(t)$. If a solution of this discrete problem exists and $\upsilon \neq 0$, the case of second order integral equations, then one gets an approximate solution for all t by

$$\upsilon x_N(t) = y(t) - q\sum_{j=1}^{N} K(t,t_j)\sigma_j$$

and therefore $x_N(t_k) = \sigma_k$. The $N \times N$ matrix

$$A_N = \|\upsilon\delta_{kj} + qK(t_k,t_j)\|_{j,k=1}^N$$

is the Nyström discretization of A. For each element of this matrix only one computation of the kernel is necessary, and often some values can be reused. This discretization is accurate for the diffraction integrals on smooth boundaries.

However, for integral operators of the first kind, $\upsilon = 0$, the simple method is not applicable, since the kernels of V, J and H are singular at the diagonal, i.e. the value of $K(t_k,t_k)$ is not defined. There exist various approaches to apply Nyström's method also to integral equations of the first kind; one of them is reported in Chapter 4. Another efficient realization was developed in Ref. 12.17 to solve boundary integral equations with the single layer potential of the Helmholtz equation, which can be easily adapted to our situation. Then again, only $\sim N^2$ computations of the kernel function are necessary and the accuracy of the modified Nyström discretization is determined by the accuracy of the computations of the kernel functions.

The situation is worse for non-smooth profiles. Then the kernels of all integrals are not differentiable at corner points and the solution of the integral equations is singular. Therefore, the usual Nyström methods with the corners among mesh points make no sense, and there exist several proposals to modify or advance Nyström methods. For example, one can form a modified smooth profile curve excluding small neighborhoods of the corner points and construct the Nyström discretization for the integrals on the modified curve. Another practice is to use graded mesh quadratures, that is, quadratures of various types which become increasingly dense near corner points. Then the discretization of the integrals on profiles with corners via the Nyström method can be made accurate; however, compared with the case of smooth profiles, the resulting matrices are excessively large and possibly very ill-conditioned, such that the method can produce inaccurate results. There are various interesting attempts in the mathematical literature to address these difficulties (see Refs. 12.18–12.20). However, it was found that

non-accounting (in respect to the usual approaches) of edge and a few other peculiarities of singular or low-convergent integral equations gives accurate and fast results for small wavelength- and height-to-period ratio diffraction problems (see Sec. 12.7.)

To solve the integral equation $Ax = y$ with a collocation method, one has to choose a set of $\tilde{N}$ approximating functions $\{\varphi_j\}$ and N collocation points $0 \le t_1 < t_2 < \ldots < t_N < 1$. The approximate solution is sought in the form

$$x_N(t) = \sum_{j=1}^{N} a_j \varphi_j(t)$$

with the unknown coefficients a_j to be determined from the collocation equations

$$Ax_N(t_k) = y(t_k), \quad \text{for all } t_k.$$

Thus the collocation discretization of A is given by the $N \times N$-matrix

$$A_N = \|A\varphi_j(t_k)\|_{j,k=1}^{N}.$$

The approximating functions should be periodic, therefore trigonometric polynomials, periodic splines (piecewise polynomials) or wavelets are good candidates. For smooth profiles we use $N = 2\tilde{N} + 1$ trigonometric monomials

$$\varphi_j(t) = \mathrm{e}^{2\pi i j t}, \quad j = -\tilde{N}, -\tilde{N}+1, \ldots, \tilde{N}-1, \tilde{N}$$

for which the expressions $A\varphi_j$ are cheap to compute. For example, the integrals with singular kernel can be written as

$$\begin{aligned} V\varphi(t) &= -\int_0^1 \log(4\sin^2\pi(t-s))\varphi(s)\,ds + \int_0^1 g_V(t,s)\varphi(s)\,ds \\ H\varphi(t) &= \int_0^1 \cot\pi(t-s)\varphi(s)\,ds + \int_0^1 g_H(t,s)\varphi(s)\,ds, \end{aligned}$$

where the functions g_V, g_H are differentiable and periodic in t and s. Thus, the integrals with these kernels can be accurately approximated with the Nyström discretization. For the main parts, the relations

$$\begin{aligned} -\int_0^1 \log(4\sin^2\pi(t-s))\varphi_j(s)\,ds &= \begin{cases} \varphi_j(t)/|j|, & j \ne 0, \\ 0, & j = 0, \end{cases} \\ \int_0^1 \cot\pi(t-s)\varphi_j(s)\,ds &= \begin{cases} i\,\mathrm{sign}(j)\varphi_j(t), & j \ne 0, \\ 0, & j = 0, \end{cases} \end{aligned} \tag{12.78}$$

are used, resulting in only one computation of the kernel functions per element of the final collocation matrix. This approach can also be used for profiles with corners, but the accuracy of the trigonometric collocation deteriorates. This is caused by the poor Fourier series approximation of functions, which are singular at corner points. In this case, collocation with splines is advantages, which are able to approximate singular functions on graded meshes. The drawback of splines collocation is that similar to the Nyström discretization graded mesh quadratures are needed to determine accurate collocation discretizations of the integrals. In Sec. 12.4.4 we describe a combination of the trigonometric and spline collocation.

12.4.3 Nyström discretization with modifications

We use the piecewise constant Nyström method (see Subsection before) with the matrix elements such as

$$A_N = \|\upsilon\delta_{ki} + qK(t_k,t_i)\|_{i,k=1}^N \tag{12.79}$$

The present Nyström method proceeds by approximating the values of the current density function φ_- of (12.43) at the quadrature nodes $(X(t_i),Y(t_i))$, $i = 1,N$; by solving the system

$$\varphi_-(X(t_i),Y(t_i)) + \sum_{k=1}^{N} c_{ik}\varphi_-(X(t_k),Y(t_k)) = b(X(t_i),Y(t_i)), \quad (X(t_i),Y(t_i)) \in \Gamma. \tag{12.80}$$

of N linear equations in the N unknowns $\varphi_-(X(t_i),Y(t_i))$ with composed coefficients c_{ik} containing matrix elements of different operators.

For relatively shallow profiles, the nodes can be uniformly put along the x-coordinates. But the approximately uniform distribution with respect to the arc-length using the parametrization (12.16) is more universal and makes it possible to treat, for example, the lamellar or any other boundary profile with abrupt slopes or non-functions by the integral method without any additional effort on the user side. The principal parameter, with respect to which the convergence is evaluated, is the number N of discretization points on each boundary. In some codes N may vary from one boundary to another, which can be useful. In the present study, let us discard this option for simplicity. Quadratures for operators with continued kernels in our codes are performed by the trapezium (= rectangle) integration rule.

The present numerical solution of the integral equations is based on a simple modification of the Nyström discretization with piecewise constant weighting functions. The choice of a discretization of integral equations such (12.80), PCGrate software default option for most cases, requires a standard regularization of integrals. For smooth curves, the convergence order is determined from the accuracy of computing the fundamental solution, which is N^{-3} if only the first derivatives of the parametrization of the curve are used. The problem becomes harder if the curve contains corners. Then the double layer potentials K and also potentials L are not compact (the single layer potential is always compact in L_2). So the usual Nyström method is very problematic to treat corners; however some interesting modifications can be found in Ref. 12.20, though they are only applicable to K and L. The treatment of the integral kernels $K(t_k,t_i)$ in (12.79) is connected with infinity for the single layer potential in the coincide points, well behaved for the double layer potential on smooth curves, and discontinuous at corner points for the double layer potential. We use three types of single-term corrections: for the single layer potential—taking into account their logarithmic singularities; for the double layer potential—accounting for the profile curvature; and for various kernel functions—acceleration terms applied to the truncated series. So, generally we do modifications in both diagonal and non-diagonal matrix elements in respect to the standard Nyström matrices having $N \times N$ regular coefficients.

The integral operators with a weak singularity in the diagonal terms can be split into a sum

$$V_{ii} = d^{-1}\int_\Gamma \ln|2\sin[(P-Q)/2]|\varphi_- d\sigma_Q + CO_{ii}, \tag{12.81}$$

where CO is a compact operator with continuous kernel. The first integral operator in (12.81) can be calculated easily for some approximations of the current density function ϕ_-. It is worth

noting that a regularization can be used even at corner nodes of a non-smooth boundary. In the presence of a profile with corners (piecewise linear), the sampling and quadrature nodes are set in such a way that all corners are nodes and the curvature corrections are applied by adding the corner term to the diagonal values. For calculations of shallow gratings having a lot of uniformly-distributed edges (multi-polygonal), another version of the quadrature formula can be applied: the nodes are set in such a way that every corner lies half-way between the nodes adjacent to it and no curvature-like single-term corrections are added (see Sec. 12.7). However, for calculations of deep grating having abrupt and/or long slopes such a simplified approach does not work and a formula involving the normal derivative of the Green function should be used.

The diagonal element of the discretization matrix corresponding to a corner point takes the value (cf. Ref. 12.21, p. 120):

$$e_{ii} = (2N)^{-1}(K^L(t_i) + K^R(t_i)) + 1/2 - \zeta_i/2\pi\,, \tag{12.82}$$

where ζ_i is the exterior angle between adjacent tangents at the corner point $P = (X(t_i), Y(t_i))$ and (cf. Ch. 4, Eq. (4.92))

$$K^{L,R}(t_i) = y'(t_i)\left[(2d)^{-1}\sum_{n=-\infty}^{\infty}\alpha_n/\beta_n + \mathrm{i}\alpha_0/2\pi\right] - y''(t_i)[(1+(y'(t_i))^2/4\pi]\,, \tag{12.83}$$

where $K^{L,R}(t_i))$ means the left- and right-sided limits, respectively, of the kernel function value at the corner point. For gratings with smooth boundaries, these corrections yield the overall error estimate $O(N^{-3})$ for diffraction amplitudes in both polarizations. However, the above simple singularity accounting is insufficient to match such accuracy of the discretization near the corners and a truncation rule together with some acceleration technique should be applied to the truncated kernel function series. Therefore, in computations of kernels, we use a direct summation approach with possible single-term corrections of corresponding matrix elements (see Sec. 12.4.5.1).

The matrices of the discretized operators contain the values of the corresponding kernel functions divided by the number N of segments between collocation points. We use in our codes a few different algorithms for solving linear systems of algebraic equations. It can be either the direct Gauss-Jordan elimination method (Gauss) or the non-direct Full orthogonalization method (FOM) which is similar to the Generalized minimum residual method (GMRES). For the FOM case, the number of iterations until a prescribed residual error is reached depends, of course, on the refraction indices, the number of accounting diffraction modes, and the profile shape, but it is nearly independent of the number of unknowns. Note that discretization of the multi-boundary integral equations can be treated by the same simple manner without modifications.

12.4.4 Hybrid trigonometric-spline collocation

Here we describe collocation methods for solving the integral equations (12.57) of conical diffraction, which contain the singular integral H^+. The collocation discretization requires the computation of this integral for special basis functions, which is simpler than the Nyström discretization, which uses point values of the strongly singular kernel function of H^+.

We consider a parametrization of Γ given by (12.16). In the case of a smooth profile Σ a trigonometric collocation method is used, i.e. we use approximations

$$\begin{aligned} w(\sigma(t))\,\mathrm{e}^{-i\alpha X(t)}\,|\sigma'(t)| &\approx w_N(t) = \sum_{k=-\tilde{N}}^{\tilde{N}} a_k\,\mathrm{e}^{2\pi ikt}\,, \\ \tau(\sigma(t))\,\mathrm{e}^{-i\alpha X(t)}\,|\sigma'(t)| &\approx \tau_N(t) = \sum_{k=-\tilde{N}}^{\tilde{N}} b_k\,\mathrm{e}^{2\pi ikt}\,, \end{aligned} \tag{12.84}$$

where the coefficients $\{a_k\},\{b_k\}$ are such that the system (12.56) is satisfied at the $N = 2\tilde{N}+1$ collocation points $t_k = k/N$, $k = 1,\ldots,N$.

The advantage of using trigonometric methods is that the integral operators $V^{\pm}$ and H^{+} with singular kernels can be approximated properly. For example, using the parametrization $\sigma(t)$ the single layer potential operator of w can be approximated by

$$V^{\pm}w(\sigma(t)) \approx -\mathrm{e}^{i\alpha X(t)}\left(\int_0^1 \log(4\sin^2\pi(t-s))\,w_N(s)\,ds + \int_0^1 g^{\pm}(t,s)w_N(s)\,ds\right),$$

and the singular integral $J^{\pm}w$ by

$$H^{+}w(\sigma(t)) \approx -\mathrm{e}^{i\alpha X(t)}\left(\int_0^1 \cot\pi(t-s)\,w_N(s)\,ds + \int_0^1 j^{+}(t,s)w_N(s)\,ds\right),$$

where the functions $g^{\pm}(t,s), j^{+}(t,s)$ are differentiable and periodic in t and s. The action of the integral operators with the kernels $\log(4\sin^2\pi(t-s))$ and $\cot\pi(t-s)$ on trigonometric polynomials is given analytically; compare (12.78). All other integrals have differentiable kernels and they are approximated by the trapezoidal rule like in the Nyström method described above. So the discretization error depends only on the error made in computing the functions $g^{\pm}(t,s), j^{+}(t,s)$ and the continuous kernels of K^{+} and L^{-}, i.e. in computing the fundamental solution and their derivatives. Here we use the exact Ewald method (cf. Section 12.4.5.2) with a number of summation terms to ensure discretization errors of order N^{-3}. Finally, the operator products $V^{+}L^{-}$, $K^{+}V^{-}$ or $H^{+}V^{-}$ are approximated by the products of the corresponding discretization matrices.

If the profile curve has corners, then the convergence properties of methods with only trigonometric trial functions deteriorate due to singularities of the densities w and τ of the form $O(\rho^{-\delta})$, $0<\delta<1$, where ρ is the distance to the closest edge. In boundary element methods it is common to use piecewise polynomial trial functions on meshes graded towards corner points. But due to the complicated form of their kernels the quadrature of the integral operators acting on piecewise polynomials is very expensive. Therefore we use a modification of the trigonometric collocation scheme with a fixed number of piecewise polynomial trial functions.

In the beginning, we introduce meshes of collocation points which contain the corners and are graded towards the corner points. This can be derived by changing the parametrization (12.16), for example, if $\sigma(t_j)$ is a corner point, then $\sigma'(t_j) = \sigma''(t_j) = 0$ implies grading towards the corner. Further, for each collocation point t_k there exists a Lagrangian trigonometric polynomial $p_k(t)$ of degree N such that

$$p_k(t_j) = \delta_{kj}\,, \quad k,j = 1,\ldots,N\,,$$

where δ_{kj} is Kronecker's delta. For each edge and a fixed number of collocation points t_k around it we replace the corresponding Lagrangian trigonometric polynomial $p_k(t)$ by a cubic

spline $s_k(t)$ on the graded mesh with $s_k(t_j) = \delta_{kj}$. Thus we get a hybrid trigonometric-spline collocation method, which combines the efficient computation of the integrals for trigonometric polynomials with the good approximation properties of piecewise polynomials on graded meshes near edges. The values at the collocation point t_j of the integrals on the basis spline s_k are computed by a composite Gauss-quadrature with a quadrature mesh geometrically graded towards t_j and depending on the distance $|\sigma(t_k) - \sigma(t_j)|$. This leads to a fixed number of additional calculations of the fundamental solutions $\Psi_{k_\pm,\alpha}$ for each discretization level compared with the pure trigonometric method, which is however compensated by a significant higher accuracy.

12.4.5 Evaluations of kernels

In spite of many research efforts (see, e.g., Refs. 12.22–12.24)—computation of the kernels remains a most time-critical part of integral method for periodic structures. Convergence of the kernels deteriorates significantly as the distance between function's arguments (a discretization point or/and a quadrature node) tends to zero, and especially near edges and at high frequencies. For more discussions we refer the reader to Ch. 4. Some "crash test" calculations on PCGrate codes can be found in Ref. 12.25 and also in numerical examples of Sec. 12.9. The Ewald sum method is quite intricate and widely used (see, e.g., Ch. 6). It is based on a separation of the infinite sum into slowly and rapidly convergent parts and, then, a transformation of the slowly convergent part using the Poisson formula and error functions. Although Ewald methods are proven to be quite efficient for many diffraction grating problems, it has turned out to exhibit poor numerical properties in short waves.

12.4.5.1 Direct kernel summation

In the described discretization method, we use a direct approach for the evaluation of kernel functions based on Poisson's summation formula (see (12.20)) and a special rule for various kernels with positive and negative summation index. In the simplest case, the series is truncated symmetrically at the lower summation index $-\tilde{P}$ and upper index $\tilde{P}$; where $\tilde{P}$ is an integer defined by

$$\tilde{P} \approx gN. \tag{12.85}$$

For many grating efficiency problems, the number of terms $\tilde{P}$ with plus or minus sign you choose should be fifty percent of a number of discretization points N ("the golden rule" and default value of PCGrate codes; for more see Sec. 12.5.7). In difficult cases like those of highly conducting blazed or very deep gratings, echelles, grazing incidence, and, especially, for the TM polarization one can try to optimize convergence and accuracy by varying g at a given number of discretization points. This parameter can be optimized at small values of N and the ratio is kept constant as N increases. Fortunately, it should be done only in very exceptional cases.

As an easy remedy to accelerate convergence of the series representing the kernels, we use the Aitken δ^2 method [12.26], which is a simple one-term improvement over a popular acceleration technique described in Ch. 4. The precision of Aitken's method for individual values of the kernels, especially at close arguments, is inferior to that provided by the Kummer acceleration used in the IESMP code ([12.27]) or by the Euler-Knopp method [12.28], but one would not benefit from extra accuracy in the end. Such more accurate acceleration techniques make sense in combination with higher-order collocation or Galerkin methods. However, it is usually

a difficult task to achieve an acceptable combination of robustness and wide-range applicability with higher-order codes. In the Aitken method suppose series $S = \Sigma a_k$ has approximately geometric convergence, then the sum

$$\tilde{S}_{K+1} = \sum_{k\le K} a_k + a_K^2/(a_K - a_{K+1})\,, \tag{12.86}$$

is a better converged series than S_{K+1} having the same number of terms. Both versions of discretization near corners together with the acceleration technique described above are found to yield approximately the same convergence rate $O(N^{-2+\varepsilon})$, where $0 < \varepsilon < 0.5$ apparently depends on boundary profile geometries.

Such a regularization of the weakly singular and singular integral operators, together with an acceleration of the truncated kernel function series, theoretically and numerically leads to higher rates of convergence and to bounded condition numbers of the discretization matrices. Though, for discretization numbers N of practical interest, no advantage of regularization is observed in our numerical experiments at very small λ/d ratios (see Sec. 12.7), even for smooth boundaries.

12.4.5.2 Ewald's method

It has turned out that acceleration techniques for the summation approach is not efficient if the second argument y has small modulus $|y|$ (cf. Ref. 12.22), which frequently occur in the quadrature of integrals for graded meshes near corners or for thin layers, i.e., very close profile curves. In this case one can use the following summation algorithm for the integral kernel which is based on Ewald's method (cf. Ref. 12.23), providing a good overview on various methods for the computation of the fundamental solution.

The idea is to split the slowly converging series (12.20) into two quickly converging series. To simplify of presentation we consider the infinite series

$$\Psi(x,y) = \frac{i}{4\pi} \sum_{n\in\mathbb{Z}} \frac{e^{inx+i\beta_n|y|}}{\beta_n} \tag{12.87}$$

with $\beta_n := \sqrt{k^2-\alpha_n^2}$ and $\alpha_n := n+\alpha$ and let $\operatorname{Re}\beta_n, \operatorname{Im}\beta_n \ge 0$. $\Psi(x,y)$ is 2π-periodic in x. Ewald's method is based on the relation

$$\frac{ie^{i\beta_n|y|}}{\beta_n} = \int_0^{a^2} \exp\left(\beta_n^2 t - \frac{y^2}{4t}\right)\frac{\mathrm{d}t}{\sqrt{\pi t}} + \frac{i}{2\beta_n}\left(e^{-iy\beta_n}\operatorname{erfc}\Big(-ia\beta_n + \frac{y}{2a}\Big) + e^{iy\beta_n}\operatorname{erfc}\Big(-ia\beta_n - \frac{y}{2a}\Big)\right),$$

which is valid for any $a > 0$ and $\beta_n \neq 0$. Here

$$\operatorname{erfc}(z) := \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-t^2}\,\mathrm{d}t$$

is the *complementary error function*. Thus we have $\Psi = \Psi^e + \Psi^w$ with the two sums

$$\Psi^e(x,y) = \frac{1}{4\pi}\sum_{n\in\mathbb{Z}} e^{inx}\int_0^{a^2} e^{\beta_n^2 t - y^2/4t}\frac{\mathrm{d}t}{\sqrt{\pi t}}\,, \tag{12.88}$$

$$\Psi^w(x,y) = \frac{i}{8\pi}\sum_{n\in\mathbb{Z}}\frac{e^{inx}}{\beta_n}\left(e^{-iy\beta_n}\operatorname{erfc}\Big(-ia\beta_n + \frac{y}{2a}\Big) + e^{iy\beta_n}\operatorname{erfc}\Big(-ia\beta_n - \frac{y}{2a}\Big)\right). \tag{12.89}$$

Since $\beta_n^2 = k^2 - \alpha_n^2$, the first sum (12.88) takes the form

$$\Psi^e(x,y) = \frac{1}{4\pi}\sum_{n\in\mathbb{Z}} e^{inx}\int_0^{a^2} e^{(k^2-\alpha_n^2)t-y^2/4t}\,\frac{\mathrm{d}t}{\sqrt{\pi t}} = \frac{1}{4\pi}\int_0^{a^2} e^{k^2t-y^2/4t}\sum_{n\in\mathbb{Z}} e^{-\alpha_n^2 t}e^{inx}\,\frac{\mathrm{d}t}{\sqrt{\pi t}}\,.$$

Poisson's summation formula gives

$$\sum_{n\in\mathbb{Z}} e^{-(\alpha+n)^2t}e^{inx} = \sqrt{\frac{\pi}{t}}e^{-i\alpha x-x^2/4t}\sum_{m\in\mathbb{Z}} e^{-\pi^2m^2/t}\,e^{\pi mx/t}\,e^{2\pi im\alpha}\,,$$

which leads to

$$\Psi^e(x,y) = \frac{e^{-i\alpha x}}{4\pi}\sum_{m\in\mathbb{Z}} e^{2\pi im\alpha}\int_0^{a^2} e^{k^2t}e^{-((x-2\pi m)^2+y^2)/4t}\,\frac{\mathrm{d}t}{t}\,. \tag{12.90}$$

Denoting $r_m^2 := (x-2\pi m)^2 + y^2$ and using the series expansion of e^{k^2t} yields

$$\int_0^{a^2} e^{k^2t}e^{-r_m^2/4t}\,\frac{\mathrm{d}t}{t} = \sum_{j=0}^{\infty}\frac{k^{2j}}{j!}\int_0^{a^2} t^{j-1}\,e^{-r_m^2/4t}\,\mathrm{d}t = \sum_{j=0}^{\infty}\frac{(ak)^{2j}}{j!}E_{j+1}\Big(\frac{r_m^2}{4a^2}\Big)$$

with the *exponential integral function* E_j of degree j

$$E_j(z) := \int_1^{\infty}\frac{e^{-zt}}{t^{\,j}}\,\mathrm{d}t\,.$$

Thus we obtain the representation

$$\Psi^e(x,y) = \frac{e^{-i\alpha x}}{4\pi}\sum_{m\in\mathbb{Z}} e^{2\pi im\alpha}\sum_{j=0}^{\infty}\frac{(ak)^{2j}}{j!}E_{j+1}\Big(\frac{r_m^2}{4a^2}\Big). \tag{12.91}$$

Since

$$E_{j+1}(z) \le \frac{\mathrm{e}^{-z}}{z+j}\,,\quad z>0\,,$$

the expression (12.91) for Ψ^e is quickly converging if $(x,y)\neq(2\pi m,0)$ with a speed of convergence increasing as the parameter a gets smaller.

The function Ψ^w can be transformed to a computationally suitable form by using the *scaled complementary error function*

$$\mathrm{w}(z) := e^{-z^2}\,\mathrm{erfc}(-iz) = e^{-z^2}\,\frac{2}{\sqrt{\pi}}\int_{-iz}^{\infty} e^{-t^2}\,\mathrm{d}t = \frac{2}{\sqrt{\pi}}\int_0^{\infty} e^{-t^2}\,e^{2izt}\,\mathrm{d}t\,, \tag{12.92}$$

which has the properties

$$\mathrm{w}(-\bar{z}) = \overline{\mathrm{w}(z)}\,,\quad \mathrm{w}(-z) = 2e^{-z^2} - \mathrm{w}(z)\,,\quad |\,\mathrm{w}(z)| \le 1 \ \text{ for } \mathrm{Im}\,z \ge 0\,. \tag{12.93}$$

Using

$$e^{\mp iy\beta_n}\mathrm{erfc}\Big(-ia\beta_n \pm \frac{y}{2a}\Big) = e^{a^2k^2}e^{-a^2\alpha_n^2}\,e^{-y^2/4a^2}\,\mathrm{w}\Big(a\beta_n \pm i\frac{y}{2a}\Big),$$

we can write (12.89) in the form

$$\Psi^w(x,y) = \frac{i\,e^{-y^2/4a^2}\,e^{a^2k^2}}{8\pi}\sum_{n\in\mathbb{Z}}\frac{e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n + i\frac{y}{2a}\Big) + \mathrm{w}\Big(a\beta_n - i\frac{y}{2a}\Big)\right). \tag{12.94}$$

From (12.93) it can be seen that $|\mathrm{w}(z)| = O(e^{(\mathrm{Im}\, z)^2-(\mathrm{Re}\, z)^2})$ if $\mathrm{Im}\, z < -|\mathrm{Re}\, z|$. To avoid numerical overflow problems, which may occur if $|y|/a$ is large, we use the relation

$$\mathrm{w}\Big(a\beta_n - i\frac{|y|}{2a}\Big) = 2\,e^{y^2/4a^2}\,e^{-a^2(k^2-\alpha_n^2)}e^{i|y|\beta_n} - \mathrm{w}\Big(-a\beta_n + i\frac{|y|}{2a}\Big) \tag{12.95}$$

obtained from (12.93), which gives

$$\frac{i\,e^{-y^2/4a^2}\,e^{a^2k^2}}{8\pi}\frac{e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n - i\frac{|y|}{2a}\Big) + \mathrm{w}\Big(-a\beta_n + i\frac{|y|}{2a}\Big)\right) = \frac{i}{4\pi}\frac{e^{i|y|\beta_n}}{\beta_n}.$$

Introducing the finite set $P := \{n \in \mathbb{Z} : \mathrm{Im}\,\beta_n + \mathrm{Re}\,\beta_n < |y|/[2a^2]\}$, the function Ψ^w is decomposed into an exponentially converging series and two finite sums

$$\begin{aligned}\Psi^w(x,y) = \frac{i\,e^{-y^2/4a^2}\,e^{a^2k^2}}{8\pi}\Bigg\{&\sum_{n\in\mathbb{Z}\setminus P}\frac{e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n + i\frac{y}{2a}\Big) + \mathrm{w}\Big(a\beta_n - i\frac{y}{2a}\Big)\right)\\ &+\sum_{n\in P}\frac{e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n + i\frac{|y|}{2a}\Big) - \mathrm{w}\Big(-a\beta_n + i\frac{|y|}{2a}\Big)\right)\Bigg\} + \frac{i}{4\pi}\sum_{n\in P}\frac{e^{inx}e^{i|y|\beta_n}}{\beta_n}.\end{aligned} \tag{12.96}$$

In particular, in the case $y = 0$ which occurs frequently for binary gratings, we obtain the exponentially converging series

$$\Psi^w(x,0) = \frac{i\,e^{a^2k^2}}{4\pi}\sum_{n\in\mathbb{Z}}\frac{e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\,\mathrm{w}\big(a\beta_n\big).$$

Note that the speed of convergence of the series in (12.96) is increasing as the parameter a gets larger.

The representation $\Psi = \Psi^e + \Psi^w$ is also used for the computation of the gradient of Ψ

$$(\partial_x + i\alpha)\Psi(x,y) = -\frac{1}{4\pi}\sum_{n\in\mathbb{Z}}\frac{\alpha_n e^{inx+i\beta_n|y|}}{\beta_n}, \quad \partial_y\Psi(x,y) = -\frac{1}{4\pi}\sum_{n\in\mathbb{Z}}\mathrm{sign}(y)e^{inx+i\beta_n|y|},$$

which is needed to compute the kernels of the operators K, L, J and H. Since $\partial_z E_j(z) = -E_{j-1}(z)$ with $E_0(z) := e^{-z}/z$, the derivatives of Ψ^e are

$$\begin{aligned}(\partial_x + i\alpha)\Psi^e(x,y) &= -\frac{e^{-i\alpha x}}{2\pi}\sum_{m\in\mathbb{Z}}(x-2\pi m)\,e^{2\pi i m\alpha}\left(\frac{e^{-r_m^2/4a^2}}{r_m^2} + \sum_{j=1}^{\infty}\frac{(ak)^{2j}}{4a^2\,j!}E_j\Big(\frac{r_m^2}{4a^2}\Big)\right),\\ \partial_y\Psi^e(x,y) &= -\frac{y\,e^{-i\alpha x}}{2\pi}\sum_{m\in\mathbb{Z}}e^{2\pi i m\alpha}\left(\frac{e^{-r_m^2/4a^2}}{r_m^2} + \sum_{j=1}^{\infty}\frac{(ak)^{2j}}{4a^2\,j!}E_j\Big(\frac{r_m^2}{4a^2}\Big)\right).\end{aligned} \tag{12.97}$$

The derivatives of Ψ^w are given by

$$\begin{aligned}(\partial_x + i\alpha)\Psi^w(x,y) = -\frac{e^{-y^2/4a^2}\,e^{a^2k^2}}{8\pi}\Bigg\{&\sum_{n\in\mathbb{Z}\setminus P}\frac{\alpha_n e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n + i\frac{y}{2a}\Big) + \mathrm{w}\Big(a\beta_n - i\frac{y}{2a}\Big)\right)\\ &+\sum_{n\in P}\frac{\alpha_n e^{inx}\,e^{-a^2\alpha_n^2}}{\beta_n}\left(\mathrm{w}\Big(a\beta_n + i\frac{|y|}{2a}\Big) - \mathrm{w}\Big(-a\beta_n + i\frac{|y|}{2a}\Big)\right)\Bigg\} - \frac{1}{4\pi}\sum_{n\in P}\frac{\alpha_n e^{inx}e^{i|y|\beta_n}}{\beta_n},\end{aligned} \tag{12.98}$$

and

$$\partial_y \Psi^w(\mathbf{x}) = \frac{e^{-y^2/4a^2}\, e^{a^2k^2}}{8\pi}\,\mathrm{sign}(y)\left\{ \sum_{n\in\mathbb{Z}\setminus P} e^{inx}\, e^{-a^2\alpha_n^2}\left(\mathrm{w}\Big(a\beta_n + i\frac{|y|}{2a}\Big) - \mathrm{w}\Big(a\beta_n - i\frac{|y|}{2a}\Big)\right)\right. \tag{12.99}$$
$$\left. + \sum_{n\in P} e^{inx}\, e^{-a^2\alpha_n^2}\left(\mathrm{w}\Big(a\beta_n + i\frac{|y|}{2a}\Big) + \mathrm{w}\Big(-a\beta_n + i\frac{|y|}{2a}\Big)\right)\right\} - \mathrm{sign}(y)\frac{1}{4\pi}\sum_{n\in P} e^{inx} e^{i|y|\beta_n},$$

where we use the relation

$$\partial_y\left(e^{-y^2/4a^2}\left(\mathrm{w}\Big(a\beta_n + i\frac{y}{2a}\Big) + \mathrm{w}\Big(a\beta_n - i\frac{y}{2a}\Big)\right)\right)$$
$$= i\beta_n e^{-y^2/4a^2}\left(\mathrm{w}\Big(a\beta_n - i\frac{y}{2a}\Big) - \mathrm{w}\Big(a\beta_n + i\frac{y}{2a}\Big)\right).$$

The numerical calculation of the exponential integral E_j and its derivatives and of the scaled complementary error function $\mathrm{w}(z)$ present no problem using standard routines. The value of the parameter a should be chosen small enough to ensure the rapid convergence of the series for Ψ^e and its derivatives and large enough to ensure the rapid convergence of the series representations for Ψ^w and its derivatives. After numerical tests we found that the choice $a|k| = 6$ is a good compromise.

12.4.6 Cache for exponential functions (plane waves)

In order to reduce time for computation matrices of the above operator equations, a simple but effective acceleration was implemented in authors' codes at the algorithmic level, i.e. cache for exponential functions (plane waves). It assumes a large time-memory tradeoff. The amount of memory required for cache can be calculated in advance in each case and adjustments (cache off or partial) are done automatically. More acceleration can be reached in some cases, e.g. if one uses cache for kernel functions (see Sec. 12.6.2). Calculation of kernel functions makes extensive use of typical multiplicative combinations of exponential functions

$$\exp\{\mathrm{i}\alpha_n(X_i - X_k) + \mathrm{i}\beta_n|Y_i - Y_k|\} = \begin{cases} E^-_{n,i}/E^+_{n,k} & \text{if } Y_i \geq Y_k; \\ E^+_{n,i}/E^-_{n,k} & \text{if } Y_i < Y_k. \end{cases} \tag{12.100}$$

Here

$$E^{\pm}_{k,i} = \exp\{\mathrm{i}\alpha_n X_i \pm \mathrm{i}\beta_n Y_i\}. \tag{12.101}$$

Let N be the number of discretization points on a given boundary, that is, the subscripts i and k in the above expressions assume N different values. Let $\tilde{P}$ be the number of exponential terms to be stored in the cache. That is, the index n assumes $\tilde{P}$ values situated symmetrically (with a possible ± 1 imbalance) with respect to 0 (see (12.85)). Normally $\tilde{P}$ is the maximum number of negative or positive exponential terms used in computations of kernel functions. If, however, there is not enough fast memory in the system, a partial cache is used, where some exponents are pre-computed and extracted form cache in the course of the kernel function computations, while other exponents are evaluated on the spot.

In total, $2\tilde{P}N$ exponents are pre-computed for every boundary. The value of P may vary, depending on which (if at all) acceleration method is used for the series summation for the

kernel functions, however for many cases $\tilde{P} \le N$. So, memory expenditure is again of the order N^2 per layer. The pre-computed exponents share the same memory for every layer, so newer values override old ones. Unlike the kernel function cache (see Sec. 12.6.2), saving the pre-computed exponents for a potential re-use in further layers with same refractive indexes does not make much sense: pre-computation only needs $O(N^2)$ operations per layer, which is a tiny fraction of the total, which is of the order N^3.

Keeping track of the stored elements order this case does not call for any special technique like as binary trees: a two-dimensional array is all one needs. However, a difficulty of another sort pops up. The numbers β_n have nonzero imaginary parts when $|n|$ exceeds some n_0, and the asymptotics of $\mathrm{Im}\,\beta_n$ is linear as $|n|$ grows indefinitely. Depending on the signs of Y_i in Eq. (12.101), the exponents easily go beyond the underflow and overflow limits in the standard floating-point arithmetic. However, the absolute values of resulting ratios (12.100) are always not greater than 1.

To resolve this problem, the data $\{E_{k,i}\}$ are stored in the format {mantissa, order}; see Ref. 12.29. The order is represented by a variable of an integer type. It can be unusually large (positive or negative) if one thinks about typical orders in engineering calculations. For example, the values $\mathrm{Im}\,\beta_n = 1000$ and $Y_i = 10$, though rather extreme, can occur in grating calculations. But for the data structure we describe, numbers like $\exp(10^4)$ are nothing unusual and totally within its capacity. In our program, the 2-byte C type `short int` is chosen for orders, which suffices for all practical purposes.

We fix a huge positive B (the "base"); in the program $B = 10^{20}$, a more or less arbitrary value. Every nonzero real or complex number Q is then uniquely represented in the form

$$Q = B^q \cdot \tilde{M}, \qquad 1 \le |\tilde{M}| < B \tag{12.102}$$

with integer q. The only arithmetical operation needed for (12.100) is division Q/Q' given that $|Q| \le |Q'|$. Assuming $Q' = B^{q'} \cdot M'$, set

$$\frac{Q}{Q'} = \begin{cases} \tilde{M}/M', & \text{if} \quad q = q', \\ (\tilde{M}/M')B^{-1}, & \text{if} \quad q = q' - 1, \\ 0, & \text{if} \quad q < q' - 1. \end{cases} \tag{12.103}$$

The divisions on the right are carried out in the in standard floating-point format.

12.5 Solving diffraction of multilayer gratings

The use of coatings has many applications in diffraction gratings. We shall consider two algorithms for conical diffraction by multilayer gratings based on the integral methods for one-profile gratings, which are theoretically able to deal with multilayered gratings without limitations concerning the shape of the interfaces or the conductivity of the layers. The choice of a numerical method to solve the multi-boundary integral equations is to a large extent independent of other implementation details of the single-boundary algorithm. It is not even necessary to use the same method for every boundary, provided that adjacent boundary solvers have a common data interface. In the hope of making the algorithm more accessible, we explicitly write out a chain of operator equations to emphasize the upper-level structure of the multilayer algorithms. Details which are not pertinent to the structural level are omitted here, but are well discussed in other Sections and Appendices. Assuming the potential operators are available

as ready-to-use building blocks, an object-oriented implementation of the operator equations becomes relatively easy.

Our description of the multilayer schemes below emphasizes its structural aspects from the perspective of an object-oriented implementation. There are two different multilayer solvers implemented in the authors' codes: the 'Separating' multi-boundary solver based on the scattering amplitude matrix algorithm described in Appendix A and the 'Penetrating' multi-boundary solver based on recursive marching algorithms described in Appendix B. The first one is restricted to multilayer gratings with horizontally separated boundary profiles, where it is possible to define a plane layer in between that does not cross the upper or the lower interface. Then one can use plane-wave Rayleigh expansions of the electromagnetic field between the interfaces and work with the scattering matrices for that interface. The second algorithm works in the case of interpenetations of interfaces, but is numerically more expensive than the first one.

Mathematical aspects of multi-boundary integral operators are nontrivial, however well represented in this Chapter and many publications. For example, transparent and detailed exposition, including a discussion of various marching schemes that avoid hypersingular potential operators, is given in Ref. 12.21.

12.5.1 Gratings with separating boundaries

Let us now consider a multilayer diffraction grating with period d formed by a stack of M relief and/or rod gratings characterized by grating profiles Σ_j, $j = 0,\dots,M-1$.

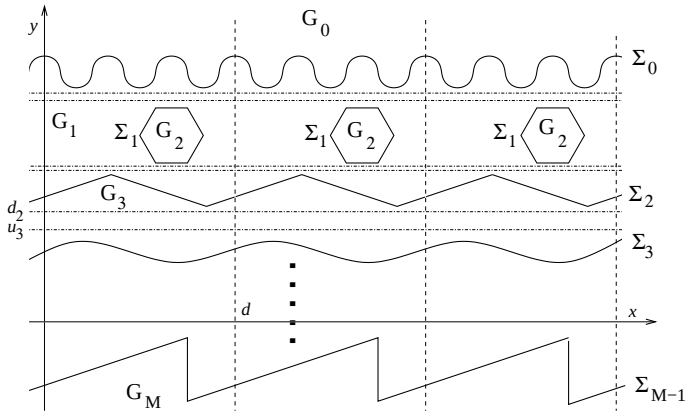


Figure 12.3: Cross section of a multilayer grating with inclusions and separating boundaries.

More precisely, the structure consists of material layers which are separated by continuous profiles and may contain rod gratings. The different media are numbered from top to bottom; see Figure 12.3, G_0 and G_M are the semi-infinite top and bottom layers. To apply a scattering matrix approach, we assume that the interfaces $\Sigma_0,\dots,\Sigma_{M-1}$ between the $M+1$ homogeneous material domains $G_0,\dots,G_M$ are separated, i.e. between adjacent interfaces Σ_j and Σ_{j-1} there exists a strip $\{u_j < y < d_{j-1}\}$ not crossing the interfaces. The structure of the multi-profile grating is characterized by the permittivity and permeability functions $\varepsilon(x,y)$ and $\mu(x,y)$, which are constant on the domains G_j. Its values in G_0 and G_M are denoted by ε_0, ε_M and μ_0, μ_M, respectively. Further we denote

$$\kappa_0^2 = \varepsilon_0\mu_0\cos^2\phi\,, \quad \kappa_M^2 = \varepsilon_M\mu_M - \varepsilon_0\mu_0\sin^2\phi\,.$$

As in the case of one interface, the z-components E_z, $B_z = (\mu_v/\varepsilon_v)^{1/2} H_z$ satisfy Helmholtz equations

$$\left(\Delta + \omega^2 \kappa^2\right) E_z = \left(\Delta + \omega^2 \kappa^2\right) B_z = 0 \tag{12.104}$$

$\kappa^2 = \varepsilon\mu - \varepsilon_0\mu_0 \sin^2\phi$, in the domains G_j and the transmission conditions at the interfaces Σ_j

$$\begin{gathered} \left[E_z\right]_\Sigma = \left[B_z\right]_\Sigma = 0\,, \\ \left[\frac{\varepsilon\,\partial_n E_z}{\varepsilon_v \kappa^2}\right]_\Sigma = -\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left[\frac{\partial_t B_z}{\kappa^2}\right]_\Sigma, \quad \left[\frac{\mu\,\partial_n B_z}{\mu_v \kappa^2}\right]_\Sigma = \sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left[\frac{\partial_t E_z}{\kappa^2}\right]_\Sigma. \end{gathered} \tag{12.105}$$

The light is incident from G_0 and we are interested in the Rayleigh coefficients $E_n^\pm, B_n^\pm$ of the series expansions

$$\begin{aligned} (E_z, B_z)(x,y) &= (E_z^i, B_z^i) + \sum_{n\in\mathbb{Z}} (E_n^+, B_n^+)\, \mathrm{e}^{i(\alpha_n x + \beta_n^{(0)} y)}, \qquad y \geq H, \\ (E_z, B_z)(x,y) &= \sum_{n\in\mathbb{Z}} (E_n^-, B_n^-)\, \mathrm{e}^{i(\alpha_n x - \beta_n^{(M)} y)}, \qquad\qquad y \leq -H, \end{aligned} \tag{12.106}$$

where the half spaces $\{y \geq H\}$ and $\{y \leq -H\}$ are contained in the semi-infinite layers G_0 and G_M, respectively. According to (12.7) we have $\beta_n^{(j)} = (\omega^2\kappa_j^2 - \alpha_n^2)^{1/2}$ if $0 \leq \arg(\varepsilon_j + \mu_j) < 2\pi$ and $\beta_n^{(j)} = -\overline{\left(\omega^2\kappa_j^2 - \alpha_n^2\right)^{1/2}}$ if $\varepsilon_j, \mu_j < 0$.

We study the off-plane diffraction for gratings with separated interfaces using the robust algorithm (for the derivation, see App. A) for modeling layered gratings (an overview is given, for example, in Ref. 12.30). The present method extends the S-matrix algorithm given by D. Maystre in Ref. 12.31 for the integral method and in-plane diffraction. It is a recursive algorithm to determine operators $\mathbf{R}_0$ and $\mathbf{T}_0$, which map the coefficients (p_z, q_z) of the incoming plane wave $(E_z^i, B_z^i) = (p_z, q_z)\, \mathrm{e}^{i(\alpha x - \beta y)}$ to the vectors of Rayleigh coefficients $\{(E_n^+, B_n^+)\}_{n\in\mathbb{Z}}$ of the reflected and $\{(E_n^-, B_n^-)\}_{n\in\mathbb{Z}}$ of the transmitted fields, cf. (12.11). To this end, the multi-profile problem is split into simpler scattering problems for one-profile gratings, which are formed by the profiles Σ_j and separate optical materials with the parameters ε_j, μ_j and $\varepsilon_{j+1}, \mu_{j+1}$.

We give a formal operator description of the marching procedure for $\mathbf{R}_0$ and $\mathbf{T}_0$. For each profile there exist scattering operators, which map the Rayleigh coefficients of an incoming field to the Rayleigh coefficients of the reflected and transmitted fields. More precisely, the grating with profile Σ_j diffracts the α-quasi-periodic incoming field

$$\sum_{n\in\mathbb{Z}} (A_n^j, C_n^j)\, \mathrm{e}^{i\alpha_n x - i\beta_n^{(j)} y}$$

in the reflected and transmitted fields

$$\sum_{n\in\mathbb{Z}} (B_n^j, D_n^j)\, \mathrm{e}^{i\alpha_n x + i\beta_n^{(j)} y} \quad \text{resp.} \quad \sum_{n\in\mathbb{Z}} (\mathscr{A}_n^j, \mathscr{C}_n^j)\, \mathrm{e}^{i\alpha_n x - i\beta_n^{(j+1)} y)}\,.$$

This is a linear operation between infinite vectors of the Rayleigh coefficients written as

$$\{(B_n^j, D_n^j)\}_{n\in\mathbb{Z}} = \mathbf{r}_j \{(A_n^j, C_n^j)\}_{n\in\mathbb{Z}}\,,\ \{(\mathscr{A}_n^j, \mathscr{C}_n^j)\}_{n\in\mathbb{Z}} = \mathbf{t}_j \{(A_n^j, C_n^j)\}_{n\in\mathbb{Z}}$$

with the linear reflection and transmission operators $\mathbf{r}_j$ and $\mathbf{t}_j$, respectively. Similarly, the reflection and transmission operators $\mathbf{r}'_j, \mathbf{t}'_j$ for illumination from below map the coefficient vector $\{(\mathscr{B}_n^j, \mathscr{D}_n^j)\}$ of the α-quasiperiodic incoming field

$$\sum_{n\in\mathbb{Z}} (\mathscr{B}_n^j, \mathscr{D}_n^j)\, \mathrm{e}^{i\alpha_n x + i\beta_n^{(j+1)} y}$$

to the coefficient vectors of the reflected and transmitted fields

$$\sum_{n\in\mathbb{Z}} (\mathscr{A}_n^j, \mathscr{C}_n^j)\, \mathrm{e}^{i\alpha_n x - i\beta_n^{(j+1)} y)} \quad \text{resp.} \quad \sum_{n\in\mathbb{Z}} (B_n^j, D_n^j)\, \mathrm{e}^{i\alpha_n x + i\beta_n^{(j)} y}\,.$$

i.e. $\{(\mathscr{A}_n^j, \mathscr{C}_n^j)\}_{n\in\mathbb{Z}} = \mathbf{r}'_j \{(\mathscr{B}_n^j, \mathscr{D}_n^j)\}_{n\in\mathbb{Z}}$ and $\{(B_n^j, D_n^j)\}_{n\in\mathbb{Z}} = \mathbf{t}'_j \{(\mathscr{B}_n^j, \mathscr{D}_n^j)\}_{n\in\mathbb{Z}}$.

Further, we assign to each profile Σ_j an y-coordinate y_j, for example $y_j = Y_j(0)$ for a given parametrisation $(X_j(t), Y_j(t))$ of the profile Σ_j, and define a diagonal operator $\boldsymbol{\gamma}_j$ which maps a vector of pairs $\{(a_n, b_n)\}_{n\in\mathbb{Z}}$ to the vector

$$\{(a_n, b_j)\, \mathrm{e}^{i\beta_n^{(j)}(y_{j-1} - y_j)}\}_{n\in\mathbb{Z}} = \boldsymbol{\gamma}_j \{(a_n, b_n)\}_{n\in\mathbb{Z}}$$

If we introduce the infinite vector $\boldsymbol{A}_0$ of the coefficients of the input wave

$$\boldsymbol{A}_0 = \{\delta_{n0}(p_z, q_z)\}_{n\in\mathbb{Z}}\,,$$

then $\mathbf{R}_0$ and $\mathbf{T}_0$ are derived by the following marching procedure:

Set	$\mathbf{R}_{M-1} = \mathbf{r}_{M-1}\,,\ \mathbf{T}_{M-1} = \mathbf{t}_{M-1}\,;$
Compute for $j = M-1, \ldots, 1$	$\mathbf{R}_{j-1} = \mathbf{r}_{j-1} + \mathbf{t}'_{j-1} \boldsymbol{\gamma}_j \mathbf{R}_j \big(\mathbf{I} - \boldsymbol{\gamma}_j \mathbf{r}'_{j-1} \boldsymbol{\gamma}_j \mathbf{R}_j\big)^{-1} \boldsymbol{\gamma}_j \mathbf{t}_{j-1}\,;$ $\mathbf{T}_{j-1} = \mathbf{T}_j \big(\mathbf{I} - \boldsymbol{\gamma}_j \mathbf{r}'_{j-1} \boldsymbol{\gamma}_j \mathbf{R}_j\big)^{-1} \boldsymbol{\gamma}_j \mathbf{t}_{j-1}\,;$
Determine finally	$\{(E_n^+, B_n^+)\}_{n\in\mathbb{Z}} = \mathbf{R}_0 \boldsymbol{A}_0\,,\ \{(E_n^-, B_n^-)\}_{n\in\mathbb{Z}} = \mathbf{T}_0 \boldsymbol{A}_0.$

12.5.2 Determination of the scattering matrices

For the application of the marching algorithm, one has to find finite-dimensional approximations of the scattering operators, i.e., scattering matrices, again denoted by $\mathbf{r}_j, \mathbf{t}_j$ and $\mathbf{r}'_j, \mathbf{t}'_j$, for given $j = 0, \ldots, M-1$. This means one-profile grating problems must be solved with incident waves from above and below for the profile Σ_j. More precisely, one has to find the Rayleigh coefficients of the diffracted fields for input waves with z-components

$$\begin{pmatrix} u_\delta^+ \\ v_\delta^+ \end{pmatrix} = \begin{pmatrix} 1-\delta \\ \delta \end{pmatrix} \mathrm{e}^{i\alpha_n x - i\beta_n^{(j)} y}\,,\ \begin{pmatrix} u_\delta^- \\ v_\delta^- \end{pmatrix} = \begin{pmatrix} 1-\delta \\ \delta \end{pmatrix} \mathrm{e}^{i\alpha_n x + i\beta_n^{(j+1)} y}\,,\ \delta = 0, 1\,. \tag{12.107}$$

The choice of the indices n will be described in Sec. 12.6.1.

First, we consider the calculation of the scattering matrices for a continuous interface Σ_j. It separates two layers and the one-profile problem corresponds to the situation depicted in Figure 12.2. We denote the semi-infinite domains above and below the profile $\Sigma = \{(x, y - y_j) : (x, y) \in \Sigma_j\}$ by $G_\pm$ and by $\varepsilon_\pm$, $\mu_\pm$ the material coefficients above and below Σ, respectively. Thus, we keep the notation of Sec. 12.2.2, but the difference to the problem there is the occurrence of different incident waves from above and below and the fixed values ε_0 and μ_0 in condition (12.105).

For illumination from above, one has to solve the following problem:
Setting

$$E_z = \begin{cases} u_+ + u_\delta^+ & \text{in} \quad G_+\,, \\ u_- & \text{in} \quad G_-\,, \end{cases} \qquad B_z(x,y) = \begin{cases} v_+ + v_\delta^+ & \text{in} \quad G_+, \\ v_- & \text{in} \quad G_-, \end{cases}$$

find α-quasi-periodic solutions of the Helmholtz equations

$$\text{in } G_+ \qquad \Delta u_+ + \omega^2 \kappa_+^2 u_+ = \Delta v_+ + \omega^2 \kappa_+^2 v_+ = 0\,, \tag{12.108}$$

$$\text{in } G_- \qquad \Delta u_- + \omega^2 \kappa_-^2 u_- = \Delta v_- + \omega^2 \kappa_-^2 v_- = 0\,, \tag{12.109}$$

where now $\kappa_\pm^2 = \varepsilon_\pm \mu_\pm - \varepsilon_0 \mu_0 \sin^2\phi$. From equation (12.105) one gets the jump conditions on Σ

$$\begin{aligned} u_- &= u_+ + u_\delta^+\,, \ \frac{\varepsilon_-\,\partial_n u_-}{\varepsilon_v \kappa_-^2} - \frac{\varepsilon_+ \partial_n (u_+ + u_\delta^+)}{\varepsilon_v \kappa_+^2} = \sqrt{\frac{\varepsilon_0 \mu_0}{\varepsilon_v \mu_v}} \sin\phi \left(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\right) \partial_t v_-\,, \\ v_- &= v_+ + v_\delta^+\,, \ \frac{\mu_- \partial_n v_-}{\mu_v \kappa_-^2} - \frac{\mu_+ \partial_n (v_+ + v_\delta^+)}{\mu_v \kappa_+^2} = -\sqrt{\frac{\varepsilon_0 \mu_0}{\varepsilon_v \mu_v}} \sin\phi \left(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\right) \partial_t u_-\,. \end{aligned} \tag{12.110}$$

For illumination from below, we set

$$E_z = \begin{cases} u_+ & \text{in} \quad G_+\,, \\ u_- + u_\delta^- & \text{in} \quad G_-\,, \end{cases} \qquad B_z = \begin{cases} v_+ & \text{in} \quad G_+, \\ v_- + v_\delta^- & \text{in} \quad G_-. \end{cases}$$

The α-quasi-periodic functions $u_\pm, v_\pm$ have to satisfy the Helmholtz equations (12.108), (12.109) and the transmission conditions

$$\begin{aligned} u_- + u_\delta^- &= u_+\,, \ \frac{\varepsilon_- \partial_n (u_- + u_\delta^-)}{\varepsilon_v \kappa_-^2} - \frac{\varepsilon_+\,\partial_n u_+}{\varepsilon_v \kappa_+^2} = \sqrt{\frac{\varepsilon_0 \mu_0}{\varepsilon_v \mu_v}} \sin\phi \left(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\right) \partial_t v_+\,, \\ v_- + v_\delta^- &= v_+\,, \ \frac{\mu_- \partial_n (v_- + v_\delta^-)}{\mu_v \kappa_-^2} - \frac{\mu_+ \partial_n v_+}{\mu_v \kappa_+^2} = -\sqrt{\frac{\varepsilon_0 \mu_0}{\varepsilon_v \mu_v}} \sin\phi \left(\frac{1}{\kappa_+^2} - \frac{1}{\kappa_-^2}\right) \partial_t u_+\,. \end{aligned} \tag{12.111}$$

Choosing as before u_-, v_- as single layer potentials (12.53), we derive from equations (12.110) and (12.111) the system of singular integral equations

$$\begin{aligned} &\left(\frac{\varepsilon_- \kappa_+^2}{\varepsilon_+ \kappa_-^2} V^+ (I - L^-) + (I + K^+) V^-\right) w - \sqrt{\frac{\varepsilon_v}{\mu_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\varepsilon_+} \sin\phi \left(1 - \frac{\kappa_+^2}{\kappa_-^2}\right) H^+ V^- \tau = U\,, \\ &\left(\frac{\mu_- \kappa_+^2}{\mu_+ \kappa_-^2} V^+ (I - L^-) + (I + K^+) V^-\right) \tau + \sqrt{\frac{\mu_v}{\varepsilon_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\mu_+} \sin\phi \left(1 - \frac{\kappa_+^2}{\kappa_-^2}\right) H^+ V^- w = V\,, \end{aligned} \tag{12.112}$$

where the singular integral H^+ is defined by (12.30) with the fundamental solution $\Psi_{\omega\kappa_+,\alpha}$. For illumination from above, the right-hand side is given by

$$U = -2\, u_\delta^+\,,\ V = -2\, v_\delta^+\,,$$

whereas in the case of illumination from below

$$\begin{aligned} U &= \frac{\varepsilon_- \kappa_+^2}{\varepsilon_+ \kappa_-^2} V^+ \partial_n u_\delta^- - (I + K^+) u_\delta^- + \sqrt{\frac{\varepsilon_v}{\mu_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\varepsilon_+} \sin\phi \left(1 - \frac{\kappa_+^2}{\kappa_-^2}\right) H^+ v_\delta^-\,, \\ V &= \frac{\mu_- \kappa_+^2}{\mu_+ \kappa_-^2} V^+ \partial_n v_\delta^- - (I + K^+) v_\delta^- - \sqrt{\frac{\mu_v}{\varepsilon_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\mu_+} \sin\phi \left(1 - \frac{\kappa_+^2}{\kappa_-^2}\right) H^+ u_\delta^-\,. \end{aligned}$$

In the case of a rod grating with a discontinuous profile, the domain G_- is bounded. Using the single layer potential ansatz in G_-, illumination from above is treated as before. Illumination from below can be treated by setting

$$E_z = \begin{cases} u_+ + u_\delta^- & \text{in} \quad G_+, \\ u_- & \text{in} \quad G_-, \end{cases} \qquad B_z = \begin{cases} v_+ + v_\delta^- & \text{in} \quad G_+, \\ v_- & \text{in} \quad G_-, \end{cases}$$

which results in the system (12.112) with the right-hand side

$$U = -2u_\delta^- \, , \; V = -2v_\delta^- \, .$$

Thus, in all considered cases the system (12.112) can be used to determine the scattering matrices. Moreover, it can be shown that the solvability of system (12.112) does not depend on ε_0 and μ_0. Similar to the system (12.56), the equations are solvable if the ratios $\varepsilon_-/\varepsilon_+$ and μ_-/μ_+ do not belong to an interval on the negative axis. Thus, the applicability of the algorithm is independent of the incidence angles θ and ϕ as well as of the polarization.

12.5.3 Gratings with penetrating boundaries

In the following, we suppose that the interfaces Σ_j are given by piecewise C^2 parametrizations

$$\sigma_j(t) = (X_j(t), Y_j(t)), \quad X_j(t+1) = X_j(t) + d, \quad Y_j(t+1) = Y_j(t)\,, \quad t \in \mathbb{R}\,, \tag{12.113}$$

i.e., the functions X_j, Y_j are piecewise C^2 with

$$|\sigma_j'(t)| = \sqrt{(X_j'(t))^2 + (Y_j'(t))^2} > 0\,.$$

Moreover, the interfaces do not intersect, i.e. $\sigma_j(t_1) = \sigma_k(t_2)$ only if $j = k$ and $t_1 - t_2 = dn$. Additionally, we suppose that if a curve Σ_j has corners, then the angles between adjacent tangents at the corners are strictly between 0 and 2π.

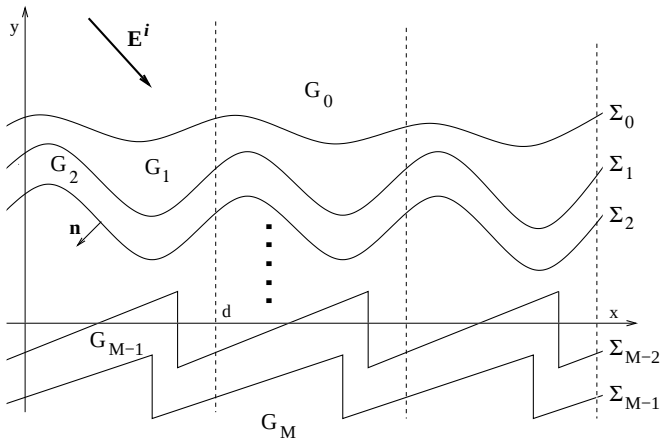


Figure 12.4: Cross section of a multilayer grating with penetrating boundaries.

To derive an integral formulation we rewrite the conical diffraction problem (12.8), (12.49), (12.11) using the notation

$$E_z(x,y) = \begin{cases} u_0 + E_z^i \\ u_j \end{cases}, \quad B_z(x,y) = \begin{cases} v_0 + B_z^i & \text{in} \quad G_0, \\ v_j & \text{in} \quad G_j\,,\; j = 1, \ldots, G_M\,, \end{cases}$$

with $E_z^i = p_z \mathrm{e}^{i(\alpha x-\beta y)}$, $B_z^i = q_z \mathrm{e}^{i(\alpha x-\beta y)}$. We seek α-quasiperiodic functions $\{u_j, v_j\}_{j=0}^N$ such that

$$\text{in } G_j \qquad \Delta u_j + \omega^2 \kappa_j^2 u_j = \Delta v_j + \omega^2 \kappa_j^2 v_j = 0\,, \tag{12.114}$$

subject to the transmission conditions

$$\text{on } \Sigma_0 \quad \left\{ \begin{array}{l} u_1 = u_0 + E_z^i\,,\ \dfrac{\varepsilon_1\,\partial_n u_1}{\varepsilon_v \kappa_1^2} - \dfrac{\varepsilon_0 \partial_n (u_0 + E_z^i)}{\varepsilon_v \kappa_0^2} = \sqrt{\dfrac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left(\dfrac{1}{\kappa_0^2} - \dfrac{1}{\kappa_1^2}\right)\partial_t v_1\,, \\ v_1 = v_0 + B_z^i\,,\ \dfrac{\mu_1 \partial_n v_1}{\mu_v \kappa_1^2} - \dfrac{\mu_0 \partial_n (v_0 + B_z^i)}{\mu_v \kappa_0^2} = -\sqrt{\dfrac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left(\dfrac{1}{\kappa_0^2} - \dfrac{1}{\kappa_1^2}\right)\partial_t u_1\,, \end{array} \right. \tag{12.115}$$

and, for $j = 1, \ldots, M-1$,

$$\text{on } \Sigma_j \quad \left\{ \begin{array}{l} u_{j+1} = u_j\,,\ \dfrac{\varepsilon_{j+1}\partial_n u_{j+1}}{\varepsilon_v \kappa_{j+1}^2} - \dfrac{\varepsilon_j\,\partial_n u_j}{\varepsilon_v \kappa_j^2} = \sqrt{\dfrac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left(\dfrac{1}{\kappa_j^2} - \dfrac{1}{\kappa_{j+1}^2}\right)\partial_t v_{j+1}\,, \\ v_{j+1} = v_j\,,\ \dfrac{\mu_{j+1}\partial_n v_{j+1}}{\mu_v \kappa_{j+1}^2} - \dfrac{\mu_j \partial_n v_j}{\mu_v \kappa_j^2} = -\sqrt{\dfrac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi \left(\dfrac{1}{\kappa_j^2} - \dfrac{1}{\kappa_{j+1}^2}\right)\partial_t u_{j+1}\,, \end{array} \right. \tag{12.116}$$

which satisfy the outgoing wave condition

$$\begin{array}{ll} (u_0, v_0)(x,y) = \displaystyle\sum_{n=-\infty}^{\infty} (E_n^{(0)}, B_n^{(0)})\, \mathrm{e}^{i(\alpha_n x + \beta_n^{(0)} y)} & \text{for } y > \max\limits_{(x,t)\in\Sigma_0} t\,, \\ (u_M, v_M)(x,y) = \displaystyle\sum_{n=-\infty}^{\infty} (E_n^{(M)}, B_n^{(M)})\, \mathrm{e}^{i(\alpha_n x - \beta_n^{(M)} y)} & \text{for } y < \min\limits_{(x,t)\in\Sigma_M} t\,. \end{array} \tag{12.117}$$

The single and double layer potentials on one period $\Gamma_j = \{\sigma_j(t) : t \in [t_0, t_0+1]\}$ of the interface Σ_j corresponding to κ_m are denoted by

$$\mathscr{S}_{\Gamma_j,m}\varphi(P) = 2\int_{\Gamma_j} \Psi_{m,\alpha}(P-Q)\,\varphi(Q)\,d\sigma_Q\,, \quad \mathscr{D}_{\Gamma_j,m}\varphi(P) = 2\int_{\Gamma_j} \varphi(Q)\partial_{n(Q)}\Psi_{m,\alpha}(P-Q)\,d\sigma_Q\,,$$

with the α-quasiperiodic fundamental solution $\Psi_{m,\alpha} = \Psi_{\omega\kappa_m,\alpha}$.

We present a recursive algorithm for solving (12.114 - 12.117), which in each step treats a problem for one of the interfaces and therefore allows us to solve conical diffraction problems for gratings with an arbitrary number of layers on standard PCs (for the derivation, see App. B). The algorithm extends a method for in-plane diffraction, i.e., $\phi = 0$, which was proposed by D. Maystre in Ref. 12.32 and described in detail in Ref. 12.4.

The starting point is to seek the solutions $\{u_j, v_j\}_{j=0}^M$ of (12.114–12.117) in the form

$$u_0 = \frac{1}{2}\big(\mathscr{S}_{\Gamma_0,0}\partial_n u_0 - \mathscr{D}_{\Gamma_0,0}u_0\big)\,, \quad v_0 = \frac{1}{2}\big(\mathscr{S}_{\Gamma_0,0}\partial_n v_0 - \mathscr{D}_{\Gamma_0,0}v_0\big)\,, \quad \text{in } G_0\,, \tag{12.118}$$

$$\left.\begin{array}{l} u_j = \dfrac{1}{2}\big(\mathscr{S}_{\Gamma_j,j}\partial_n u_j - \mathscr{D}_{\Gamma_j,j}u_j\big) + \mathscr{S}_{\Gamma_{j-1},j}\varphi_{j-1}\,, \\ v_j = \dfrac{1}{2}\big(\mathscr{S}_{\Gamma_j,j}\partial_n v_j - \mathscr{D}_{\Gamma_j,j}v_j\big) + \mathscr{S}_{\Gamma_{j-1},j}\psi_{j-1}\,, \end{array}\right\} \quad \text{in } G_j\,,\ j = 1,\ldots,M-1 \tag{12.119}$$

$$u_M = \mathscr{S}_{\Gamma_{M-1},M}\varphi_{M-1}\,, \quad v_M = \mathscr{S}_{\Gamma_{M-1},M}\psi_{M-1}\,, \qquad \text{in } G_M\,, \tag{12.120}$$

with certain densities φ_j, ψ_j, $j = 0, \ldots, M-1$. Again, the Helmholtz equations (12.50) and the outgoing wave condition (12.52) are satisfied. Note that the representations (12.118 - 12.120) are unique, provided that the single layer potential operators $V^{(j)}_{j-1\,j-1}$ are invertible for $j = 1, \ldots, M$, which will be assumed throughout.

The algorithm determines recursive relations

$$\begin{pmatrix} \varphi_j \\ \psi_j \end{pmatrix} = \mathscr{Q}_{j-1} \begin{pmatrix} \varphi_{j-1} \\ \psi_{j-1} \end{pmatrix}, \quad j = 1, \ldots, M-1, \tag{12.121}$$

such that the functions $\{u_j, v_j\}_{j=0}^{M}$ fulfill the remaining transmission conditions (12.115) and (12.116). The initial densities (φ_0, ψ_0) and the 2×2 operator matrices $\{\mathscr{Q}_{j-1}\}$ are obtained by the following scheme:

Introduce the coefficients

$$\begin{aligned} &a_j = \frac{\varepsilon_{j+1}\kappa_j^2}{\varepsilon_j \kappa_{j+1}^2}, \; b_j = \frac{\mu_{j+1}\kappa_j^2}{\mu_j \kappa_{j+1}^2}, \\ &c_j = \sqrt{\frac{\varepsilon_v}{\mu_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\varepsilon_+} \sin\phi \left(1 - \frac{\kappa_j^2}{\kappa_{j+1}^2}\right), \; d_j = \sqrt{\frac{\mu_v}{\varepsilon_v}} \frac{\sqrt{\varepsilon_0 \mu_0}}{\mu_+} \sin\phi \left(1 - \frac{\kappa_j^2}{\kappa_{j+1}^2}\right), \end{aligned} \tag{12.122}$$

and determine $\mathscr{Q}_{j-1}$ by a backward recurrence for $j = M-1, \ldots, 1$ as a solution of the operator equation

$$\left(\begin{pmatrix} I + K^{(j)}_{jj} & -c_j H^{(j)}_{jj} \\ d_j H^{(j)}_{jj} & I + K^{(j)}_{jj} \end{pmatrix} \mathscr{A}_j - \begin{pmatrix} a_j V^{(j)}_{jj} & 0 \\ 0 & b_j V^{(j)}_{jj} \end{pmatrix} \mathscr{B}_j \right) \mathscr{Q}_{j-1} = 2 \begin{pmatrix} V^{(j)}_{jj-1} & 0 \\ 0 & V^{(j)}_{jj-1} \end{pmatrix}. \tag{12.123}$$

The initial values are

$$\mathscr{A}_{M-1} = \begin{pmatrix} V^{(M)}_{M-1M-1} & 0 \\ 0 & V^{(M)}_{M-1M-1} \end{pmatrix}, \quad \mathscr{B}_{M-1} = \begin{pmatrix} L^{(M)}_{M-1M-1} - I & 0 \\ 0 & L^{(M)}_{M-1M-1} - I \end{pmatrix}, \tag{12.124}$$

and the subsequent terms in (12.123) are derived by

$$\begin{aligned} \mathscr{A}_{j-1} = &\begin{pmatrix} V^{(j)}_{j-1\,j-1} & 0 \\ 0 & V^{(j)}_{j-1\,j-1} \end{pmatrix} \\ &- \frac{1}{2} \left(\begin{pmatrix} K^{(j)}_{j-1\,j} & -c_j H^{(j)}_{j-1\,j} \\ d_j H^{(j)}_{j-1\,j} & K^{(j)}_{j-1\,j} \end{pmatrix} \mathscr{A}_j - \begin{pmatrix} a_j V^{(j)}_{j-1\,j} & 0 \\ 0 & b_j V^{(j)}_{j-1\,j} \end{pmatrix} \mathscr{B}_j \right) \mathscr{Q}_{j-1}, \end{aligned} \tag{12.125}$$

$$\mathscr{B}_{j-1} = \begin{pmatrix} V^{(j)}_{j-1\,j-1} & 0 \\ 0 & V^{(j)}_{j-1\,j-1} \end{pmatrix}^{-1} \begin{pmatrix} I + K^{(j)}_{j-1\,j-1} & 0 \\ 0 & I + K^{(j)}_{j-1\,j-1} \end{pmatrix} \mathscr{A}_{j-1} - 2 \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \tag{12.126}$$

Having found $\mathscr{A}_0$ and $\mathscr{B}_0$, the initial value (φ_0, ψ_0) of (12.121) is a solution of the linear equation

$$\left(\begin{pmatrix} I + K^{(0)}_{00} & -c_0 H^{(0)}_{00} \\ d_0 H^{(0)}_{00} & I + K^{(0)}_{00} \end{pmatrix} \mathscr{A}_0 - \begin{pmatrix} a_0 V^{(0)}_{00} & 0 \\ 0 & b_0 V^{(0)}_{00} \end{pmatrix} \mathscr{B}_0 \right) \begin{pmatrix} \varphi_0 \\ \psi_0 \end{pmatrix} = -2 \begin{pmatrix} E^i_z \\ B^i_z \end{pmatrix}. \tag{12.127}$$

Then the solution above the grating is given by the integrals

$$u_0 = -\frac{1}{2}\Big(a_0\,\mathscr{S}_{\Gamma_0,0}\big(I-L_{00}^{(1)}\big)\varphi_0 + \mathscr{D}_{\Gamma_0,0}V_{00}^{(1)}\varphi_0 + c_0\mathscr{S}_{\Gamma_0,0}J_{00}^{(1)}\psi_0\Big)\,,$$
$$v_0 = -\frac{1}{2}\Big(b_0\,\mathscr{S}_{\Gamma_0,0}\big(I-L_{00}^{(1)}\big)\psi_0 + \mathscr{D}_{\Gamma_0,0}V_{00}^{(1)}\psi_0 - d_0\,\mathscr{S}_{\Gamma_0,0}J_{00}^{(1)}\varphi_0\Big)\,.$$

If desired, the field below the grating is found from the integrals

$$u_M = \mathscr{S}_{\Gamma_{M-1},M}\varphi_{M-1}\,,\quad v_M = \mathscr{S}_{\Gamma_{M-1},M}\psi_{M-1}\,.$$

with the densities $\varphi_{M-1}, \psi_{M-1}$ determined using the recursive relations (12.121).

12.5.4 Generalization of energy balance for lossy multilayer gratings

Resonance and non-resonance anomalies, differing in their nature, can be effectively explored in high- and low- conductive gratings, such as: surface plasmon excitations, Brewster and Bragg conditions, Rayleigh orders, groove shape and waveguide features, etc. Because of the *s* and *p* modes in conical diffraction being coupled through the boundary conditions, the associated problems are more general, and gratings can act as perfect absorbers and local- or/and surface-field enhancers at any incidence polarization state.

Knowledge of the accurate value of the absorption for a grating is very important for testing the correctness and reliability of developed programs. The energy balance is one of the basic accuracy criteria based on a single computation and it is generalized here in the case of lossy multilayer gratings. In this Subsection we derive important formulas for direct calculus of the absorption of multi-boundary gratings working in general conical mounts. Diffraction efficiencies for the reflected and transmitted orders in conical diffraction can easily be found from the corresponding Raleigh coefficients or boundary values, see (12.67)–(12.70).

If the multi-boundary grating is perfectly conducting, then for respective refractive indices $\nu_j^2 = \varepsilon_j\mu_j$, $\operatorname{Im}\nu_M = \infty$, and if there is no energy absorption in the grating layers, $\operatorname{Im}\nu_j = \nu_j = 0$, $j = 1,\dots,M-1$, then the energy conservation law is expressed by the standard energy criterion (see Ch. 2) under unitary normalization conditions

$$R = 1,$$

where R is the reflected energy.

If the grating is lossless, $\operatorname{Im}\nu_j = 0$, $j = 0,\dots,M$, then the energy conservation law is expressed by a similar energy criterion

$$R + T = 1,$$

where T is the transmitted energy.

If $\operatorname{Im}\nu_j > 0$ for some $j = 1,\dots,M-1$ and $\operatorname{Im}\nu_M = 0$, then energy is absorbed in grating layers. Thus, the usual law of the energy conservation that the sum of efficiencies of all reflected and transmitted orders should be equal to the power of the incident wave, does not hold. In a general case,

$$A + R + T = 1, \tag{12.128}$$

where (see (12.157))

$$A = \frac{1}{\beta}\operatorname{Im}\int_{\Gamma_0}\Big(\frac{\varepsilon_0}{\varepsilon_v}\partial_n E_z\overline{E_z} + \frac{\mu_0}{\mu_v\beta}\partial_n B_z\overline{B_z}\Big) - \frac{\kappa_0^2}{\beta\,\kappa_M^2}\operatorname{Im}\int_{\Gamma_{M-1}}\Big(\frac{\varepsilon_M}{\varepsilon_v}\partial_n E_z\overline{E_z} + \frac{\mu_M}{\mu_v}\partial_n B_z\overline{B_z}\Big)$$

is called the absorption coefficient or simply the absorption in the given multilayer diffraction problem. If also $\operatorname{Im} \nu_M > 0$, then $T = 0$ and it holds

$$A + R = 1 \tag{12.129}$$

with the absorption (12.156)

$$A = \frac{1}{\beta} \operatorname{Im} \int_{\Gamma_0} \left(\frac{\varepsilon_0}{\varepsilon_v} \partial_n E_z \overline{E_z} + \frac{\mu_0}{\mu_v} \partial_n B_z \overline{B_z} \right).$$

The requirements (12.128), (12.129) are convenient single computation tools to check the quality of the numerical solution. Besides being physically meaningful, expression (12.157) is very useful as one of numerical accuracy tests for computational codes and especially important for x-ray–EUV gratings, photonic crystals, metamaterials, and perfect absorbers where absorption plays a predominant role. In the lossy multilayer case, one needs an independently calculated quantity A to verify (12.128). For such a quantity, we use the absorption integrals defined in Ref. 12.33 and derived in Appendix C using the second Green formula and integration by parts.

The expressions derived from the boundary integral equation theory are important for calculating the absorption of general multi-boundary gratings working in any diffraction mount at any polarization state. The boundary absorption integrals developed and tested have been found to be an accurate and universal tool for calculating of the energy balance of various periodical structures having separated or penetrating boundaries. The results of absorption calculus of a bare metallic grating with shallow grooves, photonic crystal supporting polariton-plasmon excitation and x-ray-grazing-conical-diffraction multilayer grating are demonstrated in Sec. 12.9.

Remark 12.5.1 *A generalization of the energy balance presented for a multilayer absorption grating in classical and conical diffraction is based on computations of the respective absorption integrals by values of the field and its derivatives on a boundary. It has not only intuitive significance but the same rigor, namely in the sense of generalized functions or distributions, and way to deduce as more simple energy criterions for perfectly conducting and lossless gratings (see in Ch. 2). A derivation of expressions considered for finding the absorption quantity A as well as the interpretation of the results obtained bear only on Maxwell's equations or Green's theorem and boundary conditions. The computation of A itself is not connected with a specific rigorous method which is used for near-zone field calculus. Thus, the present in-plane and off-plane energy balance generalizations for multilayer absorbing gratings can be considered as much universal and useful as well known energy conservation laws for perfectly conducting and lossless gratings.*

12.6 Implementation and algorithmic enhancements of multilayer solvers

To handle effectively various grating types, the different multilayer schemes can be used to solve respective diffraction problems, i.e. Penetrating or Separating solvers. The Penetrating solver described above is more general, since it allows the y-projections of the boundaries to be overlapping that is vital in the modeling of many coated gratings. However, when the grating boundary profiles are strictly separated, the problem (12.104)–(12.106) can be treated using certain robust algorithms for modeling layered gratings. Therefore, the Separating solver based on the S-matrix multilayer algorithm can be, for particular problems, several times faster and more accurate than the Penetrating one.

There are three basic sources of numerical errors arising from an integral equation implementation: (i) replacement of an integral equation by a finite system of linear algebraic equations; (ii) inexact evaluation of matrix elements; (iii) inaccuracy of solution of the linear system. For errors of type *i*, in many cases *a priori* estimations via functional-analytic and operator-theoretical methods are available, which, at least, can moderate one's optimistic expectations about the overall convergence rate. Combinations of moments, Galerkin, collocation, and fully discrete methods with balanced convergence properties are known as numerical schemes of discretization [12.22]. Errors of type *ii* in these methods are commonly attributed to numerical quadratures. In periodic diffraction problems, in contrast to diffraction on a compact obstacle, there is one more source of *ii*-type numerical errors: evaluation of lattice Green functions and their derivatives. The problem is seen from the well-known kernel functions representation described above. This is the most difficult error type arising from solving of grating-like diffraction problems and particularly at small λ/d ratios. Errors of type *iii*, as well as direct round-off errors, are negligible in most cases provided the numerical scheme in use is stable and the problem "generic". That is also true for iterative linear system solvers used in our codes, like GMRES- or FOM-based software (see Sec. 12.4.3).

In order to reduce time for computation matrices of the above operator equations, two further basic enhancements (cache for exponential functions (plane waves described in Sec. 12.4.6) at the algorithmic level are used in our codes: cache for kernel functions, and cash for repeating pairs or quads of layers. They assume a large time-memory trade-off. The amount of memory required for cache can be calculated in advance in each case and adjustments (cache off or partial) are done automatically. More acceleration can be reached in some cases, e.g. if one uses iterative algorithms to solve a linear system of algebraic equations (see Sec. 12.4.3).

12.6.1 Implementation of multilayer schemes

Here we describe an implementation of the S-matrix algorithm combined effectively with the conical integral equations formulated for solving such multilayer grating problems. We discuss briefly the numerical solution of the system (12.112). As mentioned before, the scattering matrices are obtained by solving one-profile equations with a finite number of illuminations (12.107). The indices n of these illuminations should be chosen such that additionally to the diffracted outgoing modes the Rayleigh coefficients of some evanescent modes are also taken into account. Let the grating formed by the profiles Σ_j, which separate optical materials with the parameters ε_j, μ_j and $\varepsilon_{j+1}, \mu_{j+1}$. The indices of propagating modes are characterized by the values n such that $\beta_n^{(j)} \geq 0$ above Σ_j and $\beta_n^{(j+1)} \geq 0$ below Σ_j. Suppose that their number is $P_u \geq 0$ above and $P_d \geq 0$ below the profile. Further, we fix numbers k_u and k_d of evanescent modes which are important to keep in the scattering matrices. This results in quadratic reflection matrices $\mathbf{r}_j$ and $\mathbf{r}'_j$ of order $2(P_u+k_u)\times 2(P_u+k_u)$ and $2(P_d+k_d)\times 2(P_d+k_d)$ for illumination from above and below, respectively. The transmission matrices $\mathbf{t}_j$ and $\mathbf{t}'_j$ are rectangular of dimension $2(P_d+k_d)\times 2(P_u+k_u)$ and $2(P_u+k_u)\times 2(P_d+k_d)$ for illumination from above and below, resp.

These matrices are constructed columnwise from the scattering amplitudes of the solutions of the equation (12.112) with right-hand sides of index n within the fixed range. Note, one has only once to discretize the integral operators in (12.112) and the LU-decomposition of this discrete matrix provides the solutions immediately and, hence, all four scattering matrices simultaneously. Therefore, we use a direct solver with LU-decomposition for computing the scattering matrices. It should be noted that modern implementations of the LAPACK and BLAS

software packages on multi- processor/core/thread machines makes direct solving a competitive alternative to iterative solution methods even for very large systems, $N \gtrsim 10000$.

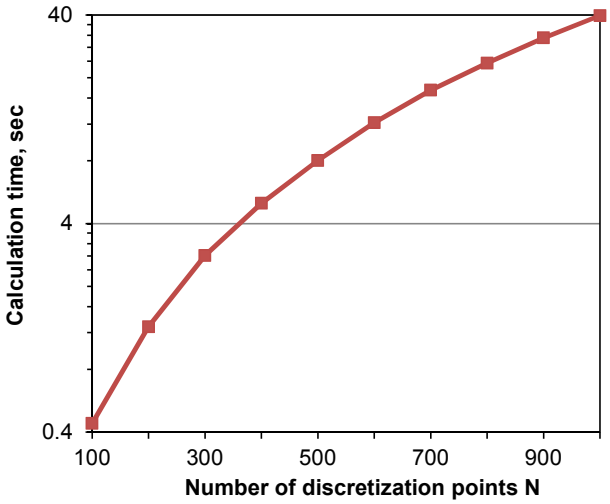


Figure 12.5: The computation time for the lamellar grating example described in Table 12.1.

Expressions (12.149) and (12.150) allow us to find amplitude matrices by a recursive procedure beginning with the lower medium. To do this, we have to know, in a general case, four matrices of scattering amplitudes and perform two matrix inversions in each iteration step. The computation time for one-boundary problems was shown to scale quadratically with the main accuracy parameter N (see Fig. 12.5). The computation time is also linearly proportional to the number of boundaries. The memory cache for amplitude matrices of multi-layer grating problems (e.g. photonic crystals) with the same boundary profiles and the same pairs or quads of layers can be used (see Section below).

Efficient implementation of the penetrating solver should use the modern implementations of the LAPACK and BLAS software packages and their analogues on multi- processor/core/thread machines. Although the algorithm requires a larger number of matrix-matrix multiplications compared to the separating solver and even the inversion of dicretization matrices, even quite complicated problems can be solved on a modern PC in reasonable time.

12.6.2 Cache for kernel functions

Matrix elements of discretized integral equations are kernel functions of one of four types considered: single-layer potentials, double-layer potentials, normal derivatives of single-layer potentials, and tangential derivatives of single-layer potentials. Any of these kernel functions for the given layer has two vector arguments: the source position $\mathbf{x}$ and the observation point $\mathbf{x}_0$. The value of the kernel function depends on the difference vector $\mathbf{d} = \mathbf{x} - \mathbf{x}_0$.

There are a number of cases of practical interest when the same difference vector $\mathbf{d}$ corresponds to more than one pair $\{\mathbf{x}, \mathbf{x}_0\}$. Typical situations include:

- conformal layers; upper and lower boundaries of such a layer are obtained from each other by a vertical shift;
- more generally, layers whose boundaries are congruent by a translation (not necessarily in the vertical direction);

- rectilinear segments of boundaries, if collocation points are uniformly distributed along such a segment.

In all these situations, it is possible to re-use values of kernel functions calculated earlier. The program stores the data: type of potential and difference of arguments vector $(\Delta x, \Delta y)$ — in lexicographical order. Fast search and insertion are provided by a binary tree structure [12.34]. Memory expenditure for the kernel function cache is of the order cN^2 per layer, where N is the maximum number of discretization points on the boundaries, and the constant c incorporates the size of data structure corresponding to each node of the tree. If no further layer has a refractive index of the current layer, then the cache gets overwritten as the solver proceeds to a new layer. However, it is quite typical to have a multilayer structure with repeating indexes, in which case the kernel function computed for one layer has a chance to be re-used in another layer. Note that the constant c is less the more effective the cache is (that is, the more repetitions occur). To save memory, single precision values are used for the difference components Δx, Δy. This approach does not compromise accuracy to any noticeable extent.

12.6.3 Cache for repeating pairs or quads of layers

The memory cache for scattering amplitude matrices (computation matrices of the operator equations considered) of multilayer grating problems with separating boundaries with the same boundary profiles and the same pairs or quads of layers can be used. The actual number of identical pairs or quads of layers can be large, up to a thousand for hard x-ray grating applications. For flexibility and possible reuse of scattering matrices of the Separating solver in multi-stack grating structures with repeated layer patterns, the dynamic caching procedure using a hash function for fast storing and extracting of boundary and adjusting layer basic parameters is initialized separately for each interface starting from the bottom. In such a procedure previously calculated instances are taken for reuse in accordance with hash function values.

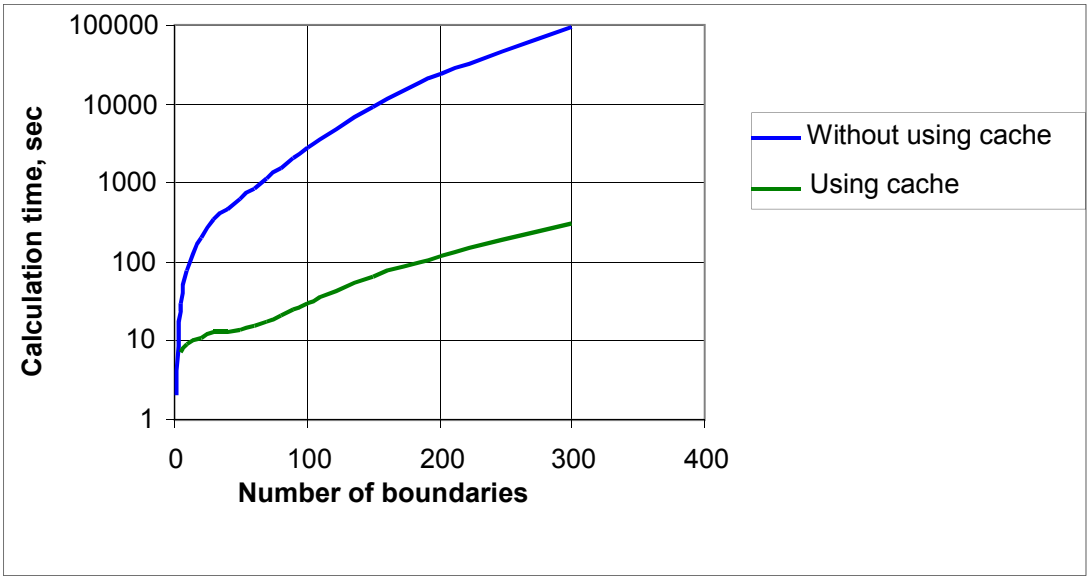


Figure 12.6: The computation time for a typical x-ray grating efficiency problem with repeated pairs of layers vs. number of identical boundaries.

Computation time of the efficiency of a coated grating with many of repeated pairs or quads of layers and equal separated boundaries can be decreased by orders of magnitude using

such memory cache for scattering matrices. For example, if the number of pairs of layer with the same boundary is more than 50, then the computation time for typical x-ray grating problems decreases more than 100 times—see Fig. 12.6.

12.7 Modifications of integral methods for very small wavelength-to-period ratios

It is well-known that solution of the 2D Helmholtz equation with any rigorous numerical code meets with difficulties at small λ/d. While the standard IMs are robust, reliable and efficient, they exhibit poor convergence and loss of accuracy in the high-frequency range due to numerical contamination in quadratures. Increasing matrix size and enhancing computation precision, as well as application of convergence speed-up techniques, which are successfully explored in low- and mid-frequency ranges, lead to unreasonably stringent requirements for computing times and storage capacities in high and, especially, ultrahigh frequency ranges ($d/\lambda > 10^3$). For various kinds of integral equations and approximation technique used for solving diffraction grating problems the computation accuracy is mostly determined from the accuracy of computing the fundamental solution (see, e.g., Sec. 12.4). In order to approximate a wave with wavenumbers $k_j = \nu_j d/\lambda$ in accordance with the Rayleigh criterion, in the Helmholtz equation one needs to use about 10 (usually from 5 to 20; it depends on a groove profile) discretization points per wavelength. So, for very large wavenumbers, say $k_j \sim 1000$, discretization matrix size should be of the order $N \sim 10^4$, a huge enough number even for modern PCs. The inaccuracy of computation of the fundamental solution, together with some rounding errors, increases significantly, up to totally divergent results, if one goes far from this simple rule of thumb. The term *modified integral method* used in publications with regard to the PCGrate software was introduced with flexibility in mind. More precisely, it is meant to be ‘"modifiable" or "tunable", however we keep the earlier term as a label. In a narrow technical sense, the MIM is characterized in this Section by a number of simple modifications required for the standard IM, similar to the one described in Ch. 4, to transform it to the MIM, together with relevant discussions.

The boundary integral equation theory is so flexible and complex that we can point out a few areas of its modifiability. (1) In the physical model, one can choose boundary types (periodical or non-periodical, closed or non-closed, smooth or having edges, randomly rough or deterministic, etc) and boundary conditions (rigorous or non-rigorous, perfect or finite conducting, extending, etc). (2) In the mathematical structure, integral representations using various potential operators and/or integral formulas together with multi-layer schemes can be considered. (3) In the method of approximation and discretization, numerical scheme of discretization (Nyström, or collocation, or Galerkin, or Method of Moments, hybrid, etc), basis and test functions (piecewise constant, or trigonometric polynomials, or splines, or delta, or Lagrange polynomials, etc), and including treatment of corners in boundary profile curves can be chosen. (4) In the low-level details, one can define methods of calculations of kernels (direct methods using Hankel or exponential functions, or Ewald’s method, or high-order summations, etc); and using acceleration techniques (Kummer or Euler-Knopp summation, or single-term corrections, etc), meshes of sampling (collocation) points (uniform or non-uniform), quadrature rules (trapezium, or Gaussian, or more sophisticated); solutions of linear systems (direct methods or iterative solvers). (5) In the implementation improvements, one can use caching of repeating quantities (exponential and kernel functions, pairs or quads of layers in multilayer structures, etc), treatment of Rayleigh orders, etc. In this Section, special attention is paid to important aspects of the presented MIM for small λ/d diffraction problems in connection with (3)–(5), and also, briefly, to some details of the numerical implementation. More about the MIM im-

plementation can be found in the documentation devoted to the PCGrate software and also in specific references [12.1].

Diffraction from 1D multilayer gratings with arbitrary boundary profile shapes, including edges, is considered in this Section in a general case of off-plane mounts. The term 1D multilayer refers to a general grating on a planar surface of arbitrary conductivity which is periodic in one direction, constant in the second, and has a finite number of borders and layers in the third. The actual number of identical or different borders and layers can be large enough, up to a few thousand for hard x-ray grating applications. Though various approximated analyses developed for the treatment of such challenging diffraction problems enjoy more or less successful application [12.35], they are always plagued with uncertainties which make comparison between rigorous and non-rigorous approaches difficult. In the present study, special attention is paid to all aspects of the MIM for $\lambda/d \ll 1$ ratios. A few commercial and non-commercial solvers based on the MIM are used in this Section and also in Sec. 12.9 to analyze the diffractive properties of various bulk and multilayer gratings including those having real boundary profiles of the polygonal type obtained by averaging measured data from Atomic Force Microscopy (AFM).

12.7.1 Approximations

As to the multilayer schemes implemented, there are no substantial differences between the well established approaches suitable for resonance domains (see Secs. 12.5 and 12.6) and the MIM in these higher levels of the multilayer boundary integral equation theory. We use both the Penetrating and Separating solvers to treat efficiencies of multilayer x-ray–EUV gratings having many boundaries with thin structure including random micro- and nano-roughness (see Sec. 12.8 and Sec. 12.9.9). However, the mid- and low-level MIM structures including the method of discretization have a few important peculiarities described below.

It is well-known the convergence and accuracy of IMs depend greatly on an appropriate choice of the discretization scheme and respective quadrature method for solving integral equation systems. As a rule, a Nyström discretization or a collocation method, as well as a Method of Moments or a Raleigh-Ritz-Galerkin approach, which are not described here, or their combination, is a good choice to treat both general and particular diffraction problems. The sampling points of unknown functions can be distributed on some uniform or multi-scale grids. In low- and mid-frequency ranges, better results are often obtained using equidistant steps along the arc length. Another possible function of the distance between collocation points is prescribed by equal steps along the x-axis perpendicular to the grooves.

In the MIM, the fastest Nyström method with the rectangular quadrature rule is used (see Sec. 12.4). Such a simple, fully discrete method combined with some matrix element modifications works well for shallow smooth boundary profiles and, particularly, at small λ/d. In the presence of a boundary profile with corners (piecewise linear), another approach can be effective. The sampling and quadrature nodes are set in such a way that every corner lies halfway between the adjacent nodes and no curvature-like single-term corrections (see (12.82)) are added to diagonal matrix elements. Let us match a solution in N midpoints $\sigma(t_{i+1/2})$ of $[\sigma(t_i), \sigma(t_{i+1})]$ by setting $\varphi_-(\sigma(t_{i+1/2})) = (\varphi_-(\sigma(t_{i+1})) + \varphi_-(\sigma(t_i)))/2$. Then we obtain a

linear system of algebraic equations for $\varphi_-(\sigma(t_k))$, $k = 1,\ldots,N$ similar to (12.80)

$$\varphi_-(\sigma(t_{i+1/2})) + \sum_{k=1}^{N} c_{i+1/2,k}\varphi_-(\sigma(t_k)) = b(\sigma(t_{i+1/2})), \quad \sigma(t_k) = (X(t_k), Y(t_k)) \in \Gamma. \tag{12.130}$$

In this approach the period of integration is divided by a number of segments equal to the number of corners on the boundary profile. Thus, the sampling points and the quadrature nodes can be put at same locations, as in (12.80), or interlacing by a half-segment shift, as in (12.130). The choice between two these complementary approaches depends on desired accuracy of computations and time requirements. For shallow boundaries with a thin structure (multi-polygonal) including roughnesses, the approach of (12.130) may have faster convergence. However, for boundaries with several rather long and abrupt slopes the approach of (12.80) may be preferable, especially if one uses (1) the curvature single-term correction by adding the corner term to diagonal matrix elements or (2) the mesh grading together with the appropriate quadrature rule, as in a case of deep gratings.

Instead of the direct summation algorithm used in the MIM and also in the IM of Ch. 4, more sophisticated methods can be implemented to accelerate the computation of the integral equation kernels, like as Ewald's methods (see Sec. 12.4.5.2). Unfortunately, it has turned out numerically that such approaches, at least those known for us, are not efficient for very small λ/d diffraction grating problems. Thus, the MIM in a narrow sense is an approximation approach with a simple discretization that also specifies a summation method for kernels.

12.7.2 Convergence and accuracy with and without speed-up technique

It is well known that the number of discretization points per wavelength used in the various IMs can be reduced significantly, up to an order of magnitude, when λ/d and H/d become small. The question is how small it might be for very small λ/d diffraction problems, say for $\lambda/d \lesssim 10^{-3}$. To accelerate convergence of the series representing kernels, different acceleration techniques can be applied (see Sec. 12.4 and also Ch. 4). In Figs. 12.7–12.10, convergence of the IM is demonstrated for an analytical case of diffraction from a plane transmission interface prescribed by a zero-depth sinusoidal profile at normal incidence in a vacuum with the lower-medium refractive index of $\nu_1 = 1.5$ and for different λ/d. Note that for all numerical examples in this Subsection, the number of positive and negative terms accounting in kernel functions was chosen at $N/2$ (see in Section 12.7.3).

For $\lambda/d = 1$ in Fig. 12.7, the convergence rate reached with speed-up techniques (all single-term corrections in kernels are used) is high, with the energy balance error of $\sim 10^{-6}$ in both polarization states for the number of discretization points $N = 10$. For $\lambda/d = 10^{-1}$ in Fig. 12.8, the convergence rate reached with speed-up techniques is medium, with the energy balance error of $\sim 10^{-5}$ in both polarizations for $N = 100$. For $\lambda/d = 10^{-2}$ in Fig. 12.9, the convergence rate, again obtained with speed-up techniques, is low, with the energy balance and transmitted energy errors of $\sim 10^{-3}$ in both polarizations for $N = 500$. The difference between the TE and TM transmitted energies for $N < 300$ is seen to be large, $\sim 10^{-1}$. For $\lambda/d = 10^{-3}$ in Fig. 12.10, the convergence rate calculated with speed-up techniques is very low, with the Energy balance error of $\sim 10^{-2}$ in both polarizations for $N = 10^3$. As seen from the figure, the convergence of the series deteriorates for $N > 1000$ as the distance between the kernel function's arguments tends to zero. In contrast to the plots of Fig. 12.10, the results for $\lambda/d = 10^{-6}$ obtained without application of any speed-up techniques exhibit the fastest

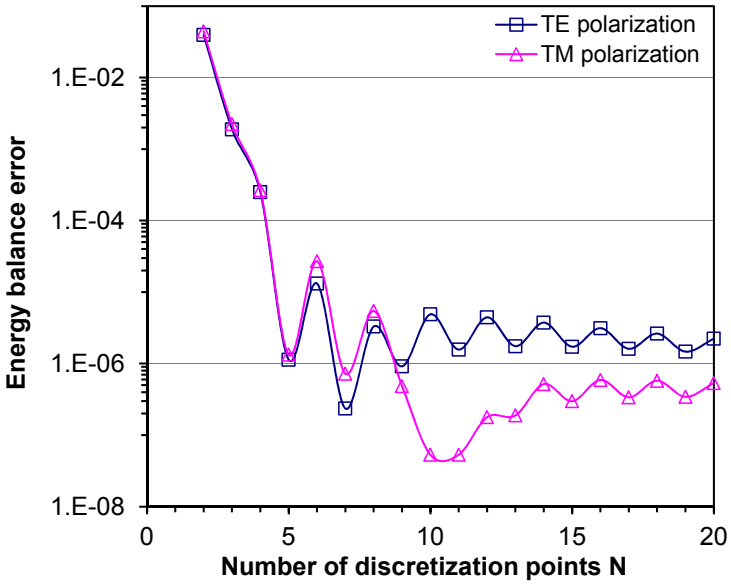


Figure 12.7: Energy balance error with the standard IM using acceleration convergence terms for the problem of diffraction on a plane transmission interface (normal incidence in vacuum with the lower medium refractive index $\nu_1 = 1.5$) vs. N for $\lambda/d = 1$.

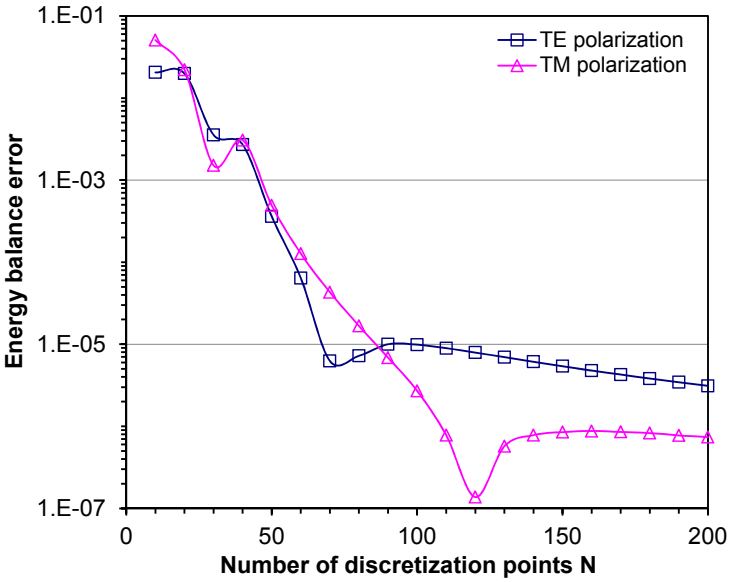


Figure 12.8: Energy errors with the standard IM vs. N used for the same diffraction problem as in Fig. 12.7 but for $\lambda/d = 0.1$.

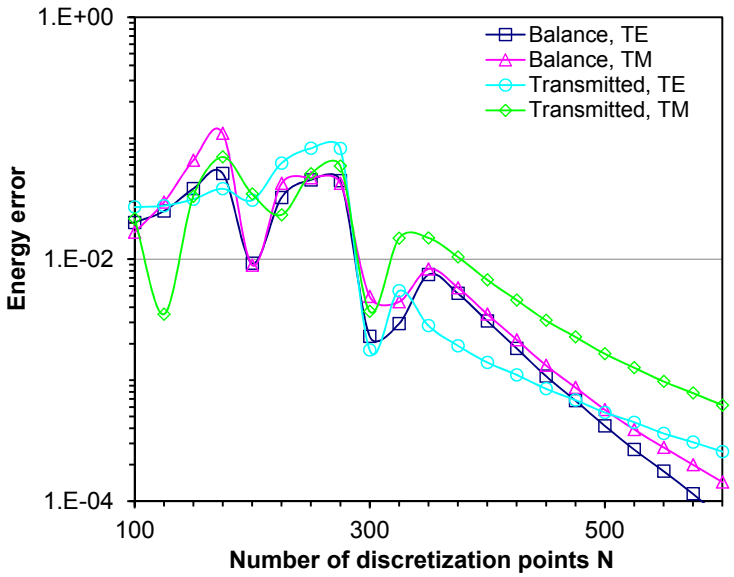


Figure 12.9: Energy errors with the standard IM vs. N used for the same diffraction problem as in Fig. 12.7 but for $\lambda/d = 0.01$.

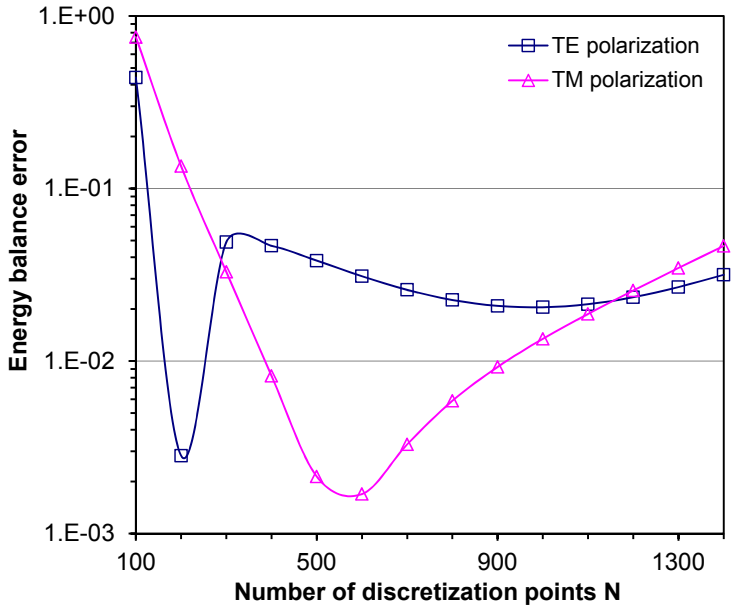


Figure 12.10: Energy errors with the standard IM vs. N used for the same diffraction problem as in Fig. 12.7 but for $\lambda/d = 0.001$.

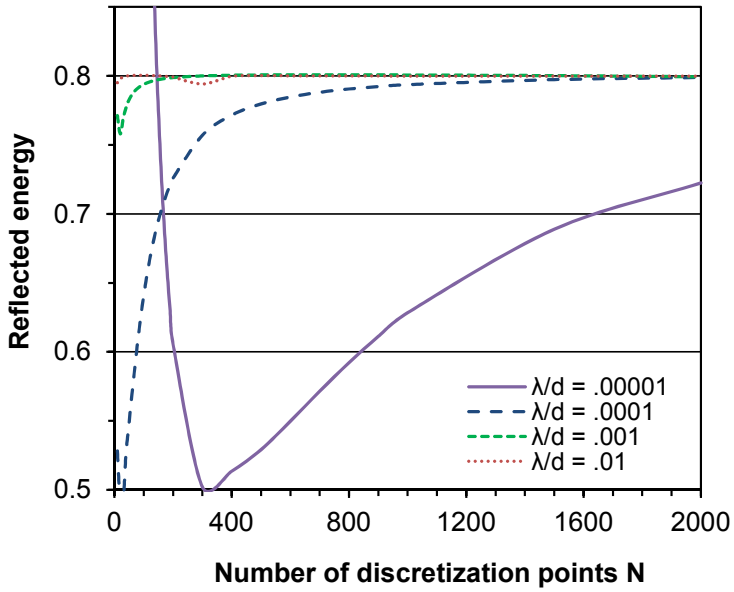


Figure 12.11: Reflected energy with the standard IM using acceleration convergence terms for the problem of diffraction on a plane Au interface of non-polarized radiation with $\lambda = 1$ nm incident at $\theta = 89°$, plotted vs. N for different λ/d.

convergence rate, with a negligible energy balance error of $\sim 10^{-16}$ for $N = 2$ only, and are equivalent to analytical calculations.

In Fig. 12.11, convergence of the standard integral method is demonstrated in respect to the main cut-off parameter N for another analytically amenable case, i.e. of x-ray diffraction from a plane absorbing interface (grazing incidence in vacuum of non-polarized radiation to plane Au surface prescribed by a zero-depth sinusoidal profile) for $\lambda = 1$ nm, $\theta = 89°$, and different λ/d. The refractive indices of Au for all examples in this Section were taken from Ref. 12.36. For $\lambda/d = 10^{-2}$, the convergence rate reached using speed-up techniques (i.e. by the standard IM) is high, with the reflected energy error of $\sim 4.9 \times 10^{-6}$ for the number of discretization points $N = 40$ (the exact reflectance value is 0.7999). For $\lambda/d = 10^{-3}$, the convergence rate reached with speed-up techniques is medium, with the reflected energy error of $\sim 10^{-3}$ for $N = 200$. For $\lambda/d = 10^{-4}$, the convergence rate, again obtained with speed-up techniques, is low, with the reflected energy error of $\sim 6.2 \times 10^{-3}$ for $N = 10^3$. For $\lambda/d = 10^{-5}$, the convergence rate calculated with speed-up techniques is very low, with the reflected energy error of $\sim 7.7 \times 10^{-2}$ for $N = 2 \times 10^3$. In contrast to the plots of Fig. 12.11, the results for extremely low λ/d of 10^{-7} obtained by the MIM without application of any speed-up techniques exhibit the fastest convergence rate with a negligible reflected energy error of $\sim 10^{-16}$ for $N = 2$ only and are equivalent to analytical calculations. Thus, we see for this grazing-incidence absorbing example the same behavior of kernel functions as in the previous absolutely different case of the normal incidence on the lossless medium.

As one can see from Figs. 12.7–12.11, at least one discretization point per wavelength is required to reach efficiency convergence for the standard IM. In contrast to that, the MIM with the simple, however very important, changes in respect to the described above standard IM, i.e., without applying acceleration convergence terms at low λ/d only, works accurately and ultra-rapidly despite the very small number of discretization points per wavelength used in the approach. For example, if a period includes $N = 10^2$ and $\lambda/d = 10^{-4}$, there is only 10^{-2} point per wavelength required for the MIM. While the results presented in Figs. 12.7–12.11

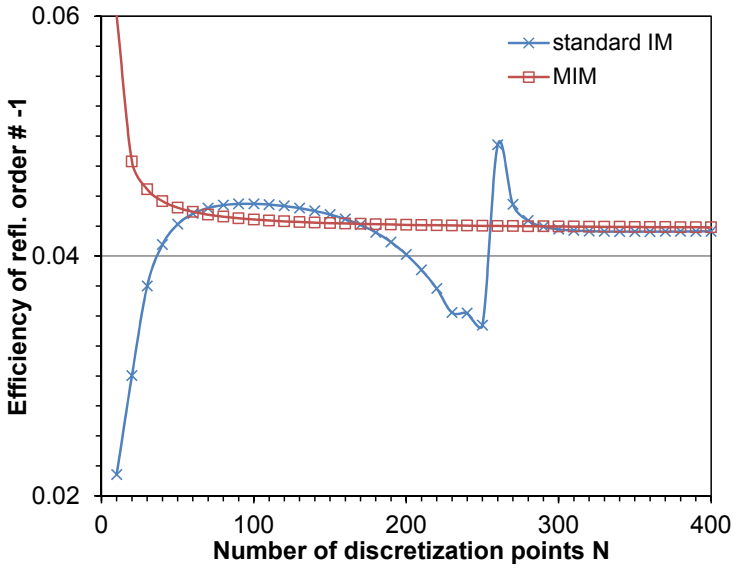


Figure 12.12: Reflected -1-order efficiency of an Au sinusoidal 300 grooves/mm grating with a depth of 25 nm illuminated by non-polarized radiation with $\lambda = 4.4$ nm incident at $\theta = 87.35°$, plotted vs. N for the standard IM or the MIM.

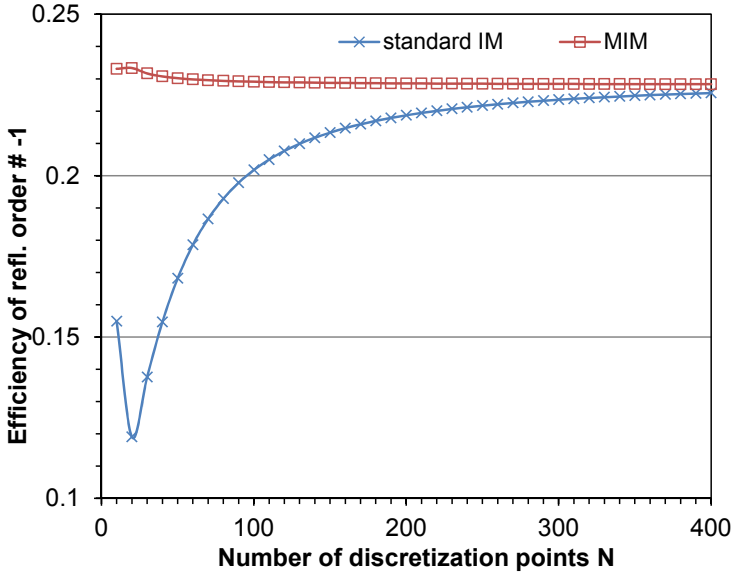


Figure 12.13: Reflected -1-order efficiency of an Au sinusoidal 3600 grooves/mm grating with a depth of 10.5 nm illuminated by non-polarized radiation with $\lambda = 4.4$ nm incident at $\theta = 86.15°$, plotted vs. N for the standard IM or the MIM.

may certainly be different for various realizations of the IMs and of the speed-up techniques used, the overall pattern remains the same.

Shallow gratings and rough mirrors exhibit a similar behavior for very small λ or λ/d (very large wavenumbers k) in the x-ray and EUV ranges (see Figs. 12.12 and 12.13). In this case, however, the effective boundary profile depth $\sim H\cos\theta$, the bilayer thickness $\sim \lambda/(2\cos\theta')$ ($\theta' = \arcsin[(\sin^2\theta\cos^2\phi + \sin^2\phi)^{0.5}]$), and the effective radiation wavelength $\sim \lambda/(\nu_j\cos\phi_j)$ must be of the same order of magnitude. In the present approach, the peculiarity described in Ref. 20 ("Introducing known speed-up terms in integral methods produces an adverse numerical effect because of the ensuing uncontrolled growth of coefficients in analytically (or numerically—Goray & Schmidt) improved asymptotic estimations") takes into account mostly for the case of shallow x-ray–EUV gratings working at very small λ/d and including, if any, random roughness (for more, see Remark below and also Section 12.8).

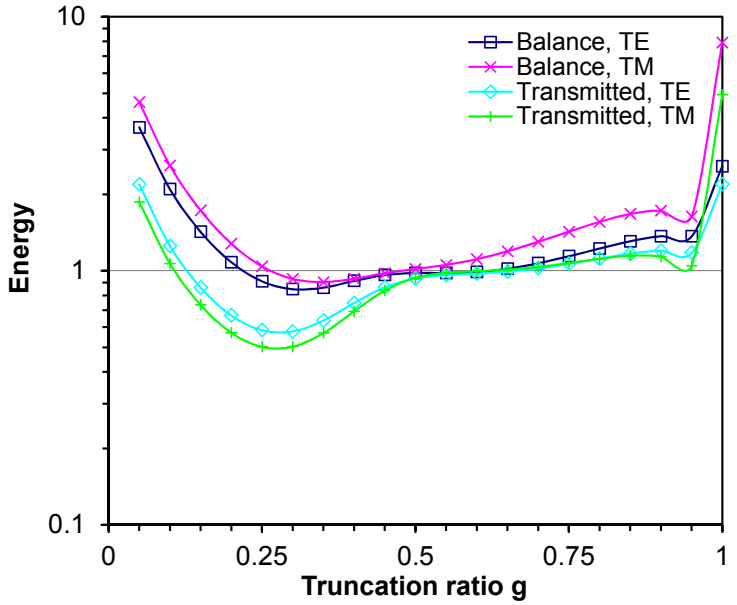


Figure 12.14: Energy balance and transmitted energy with the MIM, plotted vs. g at $N = 100$ for the same diffraction problem as in Fig. 12.7 but for $\lambda/d = 0.01$.

As shown in this Subsection *with all speed-up options turned off, it is possible to obtain for the most difficult problems of small λ/d ratios surprisingly rapid convergence, and an energy balance very close to* 1. The most important among the convergence speed-up options which have to be switched off in this case is the allowance for logarithmic singularity, and second in importance is the correction applied to account for the cut-off terms in the expansions of kernels (the Aitken's δ^2 single-term correction in our case (see Sec. 12.4.5.1)). Switching off the curvature single-term correction is of lower but not minor significance on the way to reaching fast convergence. Such calculations at very low λ/d also depend significantly on the actual summation rule chosen for the kernel functions that will be discussed in next Subsection.

Remark 12.7.1 *The same rule for the relations between basic grating and light parameters and reaching the maximum diffraction efficiency in a desired order is, on the whole, valid for longer wavelengths, too. For example, the MIM with speed-up options turned off can be applied also for echelle gratings working at very high orders (very low λ/d) and $H\cos\theta/d \ll 1$ [12.5, 12.35]. Thus, the record of rigorous computations was achieved for the r-10 EXES echelle grating of the NASA project SOFIA with $d = 7.62$ mm working in the -1431 order at a wavelength*

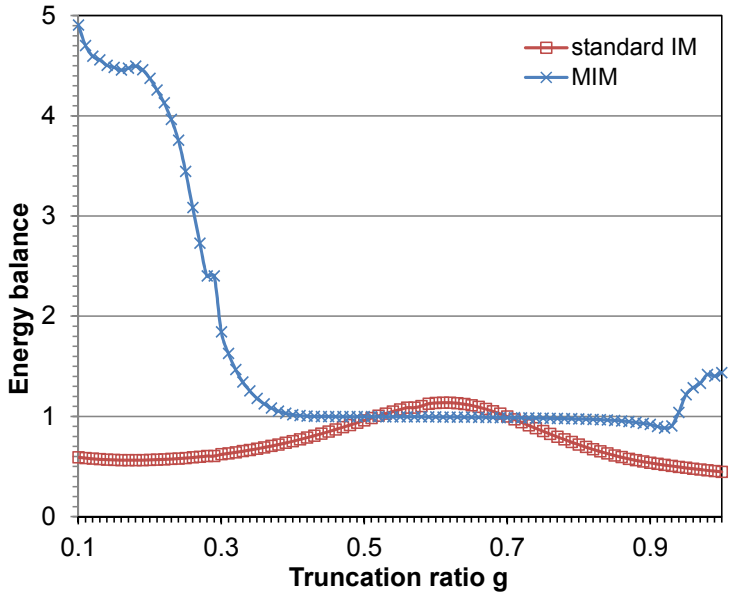


Figure 12.15: Energy balance of an Au sinusoidal 300 *grooves/mm grating with a depth of* 10.5 *nm illuminated by non-polarized radiation with* $\lambda = 0.834$ *nm incident at* $\theta = 88.65°$*, plotted vs. g for* $N = 100$ *and the standard IM or the MIM.*

of 10.6-μ*m [12.1]. Because of the very small wavelength-to-period ratio* ($\lambda/d \sim 0.001$) *and rather deep profile depth* ($2H/d \sim 0.1$) *it is necessary for such a case to increase the truncation parameter N up to a value of* ~ 3000.

12.7.3 Summation rules for kernel functions and energy balance

For many practical cases, there are no big problems to reach fast convergence and sufficient accuracy of results obtained by only varying the major accuracy parameter N_j, which is usually the same for all boundaries of gratings layers: $N_j = N$. The N values of 100–400 provide good accuracy commonly, with the exception of the following difficult cases: very deep (in respect to period and/or wavelength) boundaries; real boundary profiles with super fine structures including random roughness; very close boundaries; extremely grazing incidence; bad points on Rayleigh wavelengths, resonance anomalies of different kinds; high order echelles; high conductivity (especially for the TM polarization); some others and, especially, a combination of a few of these cases. For such complex problems, an increase in the number of discretization points may become necessary. However, to obtain accurate data for hard examples of computations, i.e. at very low λ/d, optimization of another accuracy parameter should be fulfilled.

In addition to N, there is one more important code parameter, namely the "Maximal number of accountable plus or minus terms" that describes a number of positive and negative terms accounting in kernel functions. This is the number of grating adjacent periods accountable in expansions of Green functions and their derivatives due to the quasi-periodicity property of the fields. In the simplest case typical of real problems, all kernels are truncated symmetrically in respect to the upper and lower regions and equally for any j-th boundary:

$$\tilde{P}_j^{\pm} = \tilde{P}_j = \tilde{P} \approx gN_j = gN. \tag{12.131}$$

The "truncation ratio" g is optimized at small values of N and is kept constant as N increases. It has been found [12.25] that $g = 1/2$ is a reasonably good choice for most practical compu-

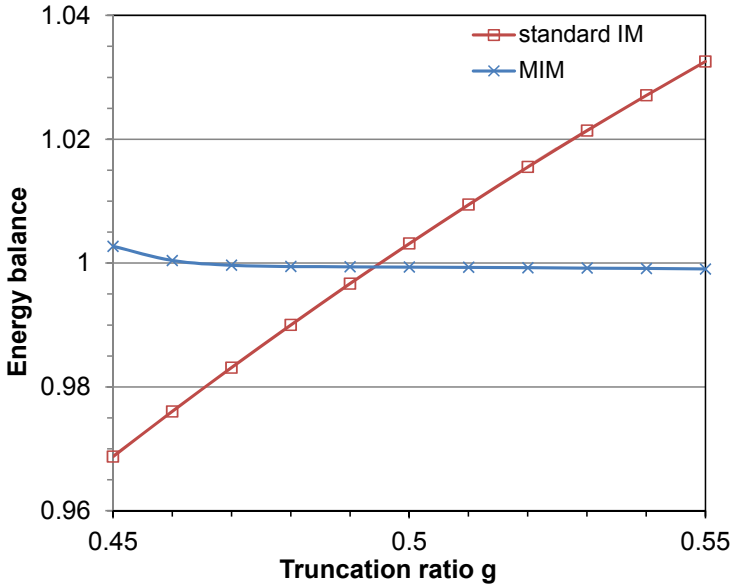


Figure 12.16: Energy balance vs. g used for the same diffraction problem as in Fig. 12.15, but for $N = 400$.

tations, and especially in the short wavelength range. Typical dependencies on $\tilde{P}$ for the above example with $\lambda/d = 10^{-2}$ are shown in Fig. 12.14. The energy balance is closer to 1 in both polarizations and TE/TM transmitted energies are close to each other at $\tilde{P} = 50\%$ of N, with divergence seen to set in at smaller and larger values of g.

For another, very different, example of the absorbing x-ray grating working at grazing incidence, one can see similar dependencies of the energy balance on $\tilde{P}$ in Figs. 12.15 and 12.16 for different numbers of discretisation points. The energy balance is close to 1 for both integral methods considered at $P \approx 50\%$ of $N = 10^2$ with a very high rate of convergence for the MIM and very slow convergence for the standard IM, similar to the convergence dependencies on N presented above. While the MIM has the long-range of high accuracy converged results from $\tilde{P} \approx 40\%$ to $\tilde{P} \approx 70\%$ in Fig. 12.15, only two points in a curve for the standard IM have the energy balance values close to 1, with divergence seen at both sides from these points. Similar behavior is seen in Fig. 12.16 for the high value of $N = 400$, where again the energy balance is close to 1 for the MIM and the standard IM at $g \approx 0.5$.

While today this rule is no more than empirical, there can be no doubt whatsoever that this choice is valid, and this has been verified in many realistic examples during the past two decades. Note that in the IM developed by D. Maystre during the later 70s [12.37], $g = 2/3$ for the resonance domain ($\lambda \sim d$) and should be varied for different λ/d . It is worth noting that $g = 2/3$ is worse than $g = 1/2$ because the computation time is proportional to $2\tilde{P}N^2$. It is interesting that the first "good" point in Fig. 12.15 for the standard IM is close to the value of $g = 0.5$, i.e. our "golden rule", and the second "good" point—to the value of $g = 0.7$, which agrees well with the rule of $g = 2/3$ given earlier for the standard IM. The present golden rule is also approximately satisfied for the all examples of numerical results given in Ch. 4.

To reduce computing time for matrices of the discretized operator equations, a few enhancements at the algorithmic level are used in the MIM: cache for kernel functions, cache for exponential functions, and cache for repeating pairs or quads of layers of multilayer gratings (see Secs. 12.4.6, 12.6.2, and 12.6.3). They assume a big time-memory trade-off at small λ/d. The amount of memory required for cache can be calculated in advance in each case and ad-

justments (cache off or partial) are done automatically. The computation algorithm group in PCGrate codes enables one also to choose an algorithm for solving linear systems of algebraic equations. It can be either the direct Gauss or the non-direct FOM method (see Sec. 12.4.3). Note that in the Penetrating solver, linear systems are solved with the Gauss algorithm alone.

Remark 12.7.2 *One more important note regarding the energy balance summation for very small λ/d problems appears to be pertinent here. The Green function and its derivative members tend to big values near Raleigh wavelengths when the y-component $\beta_n^{(0),(M)}$ of the n-th diffraction order wave vectors in the upper medium or/and in the lower medium (for transmission gratings) tends to zero (see* (12.20)*). This means that the diffraction order becomes grazing or even close to evanescent. Its efficiency may be rather high from the physical point of view or/and diverge from the mathematical point of view (it depends also on N). It is well known from the diffraction theory that the efficiency of strictly grazing propagating, as well as of all evanescent, orders is zero. Moreover, various rigorous and approximate methods valid for shallow gratings operating at small λ/d, as well as all experimental data suggest convincingly that the efficiency decreases rapidly with increasing modulus of the diffraction order number. As a rule, the efficiencies of such grazing orders should be very close to zero and much less than the inaccuracy of computations. Thus, rather big and diverging efficiencies of high number grazing orders should be excluded from the energy balance consideration, for example, starting from a high order which becomes increasing in efficiency.*

12.8 Analysis of rough gratings using quasi-periodicity and Monte Carlo calculus

Multi-wave and multiple diffraction, refraction, absorption, waveguiding and wave deformation govern, to a considerable extent, scattering of x-ray and EUV radiation and cold neutrons from nanoroughness of continuous media. Inclusion of these pure dynamic effects, which requires application of electromagnetic theory, permits one to calculate the absolute intensity of coherent (specular or diffraction order) components and describe adequately the intensity distribution of the non-coherent (diffuse) components which may have resonance peaks. Some surfaces are deterministic, e.g., perfect gratings, and some are random, e.g., polished mirrors). Some surfaces are 1D, e.g., one-periodic (classical) gratings and cutting mirrors, but most are 2D, e.g., bi-periodic gratings (bigratings), ocean surfaces, and surfaces with atomic scale roughness. Any number of possible combinations between these four characteristics may be present in real structures, e.g., 1D deterministic gratings modulated with 2D random roughness. Despite the impressive progress reached recently in development of exact numerical methods of investigating wave diffraction from boundary roughness [12.38, 12.39], the present authors are aware only of asymptotic and perturbation approaches to the analysis of x-ray and neutron scattering for 1D and 2D rough surfaces, such as the scalar Kirchhoff integral, parabolic wave equation, Rayleigh method, Born approximation, distorted-wave Born approximation, and a few others [12.40, 12.41]. The MIM and other rigorous approaches identified that the intensities of x-ray–EUV scattering at boundaries with random roughnesses may differ considerably (by a few times) from the values derived with the use of various approximate models [12.5, 12.6]. It was found that the MIM operates equally well with nano-roughness of any kind and shape which obey arbitrary statistics of distribution (not necessarily periodic or Gaussian, or fractal, etc.).

There are two classical and equivalent approaches, with some restrictions in each of them, to model rigorously randomly-rough 1D and 2D surfaces. The most general and time-consuming one is to use large surface lengths of many wavelengths. In this approach some

window functions and tapered (narrowing) beams can be used to restrict the illuminated range and avoid numerical difficulties at endpoints. The second widely-explored approach is to use periodic boundary conditions (quasi-periodicity of Floquet-Bloch modes). This method uses an infinite beam (plane wave) and assumes that the random rough surface lengths repeats itself for given large periods having some numbers of random asperities. That means using infinite grating samples together with intensive Monte-Carlo simulations. Examples of the both famous approaches are well described in the literature, see e.g. in Refs. 12.38, 12.42–12.44. From the theoretical and numerical reasons we thought it convenient to use the large-period-grating model to analyze shallow randomly-rough gratings in the x-ray–EUV range. This classical model for computation of bulk or few-border rough mirrors and quasi-gratings is applied in PCGrates and other of our codes to calculate multilayer rough mirrors and gratings, as well as multiple quantum dot or quantum molecular ensembles with most realistic border profiles having irregularities of any kinds, including real ones, i.e. measured by AFM, Transmission Electron Microscopy, Near-field Scanning Optical Microscopy, etc, or derived from simulations using a continuum growth model of multi-scale reliefs [12.45–12.47].

Diffraction from 1D surface grating-like structures with shallow boundary profile shapes is considered in this Section for the sake of simplicity for bulk gratings working in conical diffraction at small λ/d ratios. A generalization to a multilayer case is straightforward. The integral equations developed in the previous Sections are used in the present Section to analyze the diffractive properties of bulk gratings with real-profile boundaries having random roughnesses. The Section also reports on the electromagnetic solution of reflection from 2D rough surfaces in short waves using boundary integral equations for gratings in conical diffraction and Monte Carlo simulations. The general equivalence rule for determination of the efficiencies of reflected orders of bigratings (2D) from those calculated for classical (1D) gratings is derived.

12.8.1 Scattering intensity, absorption, and energy balance of rough 1D gratings

For a given incident plane wave with wave vector

$$\mathbf{k} = (\alpha, -\beta, \gamma) = k_+(\sin\theta\cos\phi, -\cos\theta\cos\phi, \sin\phi),$$

the reflected and transmitted diffraction orders of number n have the wave vectors

$$\mathbf{k}_n^{\pm} = (\alpha_n, \pm\beta_n^{\pm}, \gamma) = k_{\pm}(\sin\theta_n^{\pm}\cos\phi^{\pm}, \pm\cos\theta_n^{\pm}\cos\phi^{\pm}, \sin\phi^{\pm}),$$

with $(k^{\pm})^2 - \gamma^2 = {\alpha_n}^2 + (\beta_n^{\pm})^2$, $(\beta_n^{\pm})^2 \geq 0$. Since the z-dependence of all functions is given by $\exp(i\gamma z)$

$$\tan\theta_n^{\pm} = \alpha_n/\beta_n^{\pm}, \quad \phi^{+} = -\phi, \quad \phi^{-} = \arcsin(k_+\sin\phi/k_-).$$

By convention, the outgoing angles $\theta_n^{\pm}$ of the reflected and transmitted orders (to ensure that $\theta_0^{+} = -\theta$) are taken from the interval $[-\pi/2, \pi/2]$, as well as ϕ^{+} and ϕ^{-}.

The p- and s-components of the E-fields of the plane waves (incident and diffracted) are defined with respect to the grating normal $n = (0,1,0)$. We define the vectors $\mathbf{s}$ orthogonal to the plane spanned by $\mathbf{k}$ and the grating normal and $\mathbf{p}$ lying in that plane (see Sec. 12.2):

$$\mathbf{s} = \mathbf{k}\times(0,1,0)/|\mathbf{k}\times(0,1,0)|, \ \mathbf{p} = \mathbf{s}\times\mathbf{k}/|\mathbf{k}|.$$

If $\mathbf{k} = (0,k,0)$, we set $\mathbf{s} = (0,0,1)$ and hence $\mathbf{p} = (1,0,0)$. Since for a plane wave the electric field $\mathbf{E}$ is orthogonal to the wave vector, $(\mathbf{E},\mathbf{k}) = 0$, one can decompose $\mathbf{E}$

$$\mathbf{E} = (\mathbf{E},\mathbf{s})\,\mathbf{s} + (\mathbf{E},\mathbf{p})\,\mathbf{p}.$$

The polarization angles of the wave are defined as

$$\delta = \arctan[|(\mathbf{E},\mathbf{s})|/|(\mathbf{E},\mathbf{p})|],$$
$$\psi = -\arg[(\mathbf{E},\mathbf{s})/(\mathbf{E},\mathbf{p})],$$

where $\delta \in [0,\pi/2]$, $\psi \in (-\pi,\pi]$. Such a representation of polarization angles is useful to define polarization states and polarization properties of incoming and diffracted waves in diffraction grating applications, see, e.g., Examples in Sec. 12.9.

The efficiency of a diffracted order represents the proportion of power radiated in each order. Defining the power as the flux of the Pointing vector modulus $|\mathbf{P}| = \mathrm{Re}\,(\mathbf{E}\times\overline{\mathbf{H}})/2$ through a normalized rectangle parallel to the (x,z)-plane, the ratio of the power of a reflected or transmitted propagating order and of the incident wave gives the conical diffraction efficiency $\eta_n^{\pm}$ of this order in the simple form (see (12.67), (12.68)). For the reflected orders we have

$$\eta_n^+ = \frac{\beta_n^+}{\beta}\left(\frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2 + \frac{\mu_+}{\mu_v}|B_n^+|^2\right),$$

where the formulas for $E_n^{\pm}$, $B_n^{\pm}$ are given by (12.69) and (12.70). If $\mathrm{Im}\,k^- > 0$ then there are no transmitted orders. Thus, the usual law of energy conservation, that the sum of efficiencies of all reflected and transmitted orders should be equal to the power of the incident wave, does not hold. Instead, some part of the power is absorbed in the substrate. If the grating is absorbing, then conservation of energy is expressed by a criterion

$$R + A = \sum_{\beta_n^+ > 0} \eta_n^+ + A = 1, \tag{12.132}$$

where R is the sum of the reflection order efficiencies and A is the absorption in the single-boundary off-plane problem that can be computed from integrals of the solution of the partial differential formulation of conical diffraction (see (12.74)). For the general elliptically polarized incident light in conical diffraction, the reflected efficiency can be found as

$$\eta_n^+ = |C_n^+(\theta,\phi,\delta,\psi)|^2 \beta_n^+(\theta_n^+,\phi^+)/\beta(\theta,\phi), \tag{12.133}$$

where $|C_n^+|^2$ for a reflected order of the number n in conical diffraction is expressed by

$$|C_n^+|^2 = \frac{\varepsilon_+}{\varepsilon_v}|E_n^+|^2 + \frac{\mu_+}{\mu_v}|B_n^+|^2.$$

As mentioned in Sec. 12.3.2, the balance requirement (12.132) is one of the most important accuracy criteria based on a single computation generalized in the lossy case by the explicit computation of A from (12.73). The sum $R+A$ is actually the energy balance for an absorbing grating in conical diffraction, including that having rough grooves, and the extent to which it approaches unity is a measure of the accuracy of a calculation.

For $\lambda/d \ll 1$ the discrete order efficiencies is an approximation of the differential reflection coefficient (DRC) ς (analogous of a bistatic scattering coefficient [12.38]) for a continuum of scattered angles so that

$$\sum_{\beta_n^+ > 0} \eta_n^+ = \int_{-\pi/2}^{\pi/2} \varsigma(\theta_n^+)\, d\theta_n^+. \tag{12.134}$$

From the grating equation for conical diffraction in the form

$$k_{xn}^{+} = k_x + \frac{2\pi n}{d}, \tag{12.135}$$

where $k_{xn}^{+} = k\sin\theta_n^{+}\cos\phi$ and $k_x = k\sin\theta\cos\phi$ we know the derivative

$$\frac{dk_{xn}^{+}}{dn} = \frac{2\pi}{d}. \tag{12.136}$$

Then, for large enough N, $|n| \le N$ and accounting $dn = 1$ one can derive

$$\sum_{\beta_n^{+}>0} dn = \sum_{\beta_n^{+}>0} = \frac{d}{2\pi}\int_{-\pi/2}^{\pi/2} k\cos\theta_n^{+}\cos\phi\, d\theta_n^{+}. \tag{12.137}$$

From (12.133), (12.132), and (12.137) we have

$$\sum_{\beta_n^{+}>0} |C_n^{+}|^2\beta_n^{+}/\beta = \frac{d}{2\pi}\int_{-\pi/2}^{\pi/2} \frac{k\cos^2\theta_n^{+}|C_n^{+}|^2\cos\phi}{\cos\theta} d\theta_n^{+}. \tag{12.138}$$

Compare (12.134) and (12.138) we obtain the DRC for conical diffraction

$$\varsigma(\theta_n^{+}) = \frac{d\cos^2\theta_n^{+}|C_n^{+}|^2\cos\phi}{\lambda\cos\theta}. \tag{12.139}$$

The general case of 2D rough surfaces may be considered in a similar way. It can be done, for example, by expressing the solution of the 3D Maxwell equations for bigratings through solutions of the 2D Helmholtz equation for classical gratings working in conical diffraction, an approach which may be resorted to in some important cases (see Sec. 12.8.3).

12.8.2 Scattering intensity of rough gratings in a dispersive plane

For accounting random roughness rigorously in our codes, we use the model in which the randomly rough surface is represented by a grating of large period d. This period may contain a few or a large number of random asperities and/or a discrete number of periodic grooves. So the program deals with a structure that is a grating from a mathematical point of view but that can model a randomly rough surface of a grating or a mirror. If the groove spacing becomes small compared with the correlation length ξ of the random asperities, then the discrete dimension scaling can be applied to such a rough grating and the diffraction is investigated on the equivalent surface structure in proportionally higher diffraction orders. Moreover, if the width of the asperities has the same order of magnitude as the wavelength of incident light, the number of diffraction order is large, and the continuous speckle of the randomly rough surface is simulated by the discrete speckle of the grating, as has been demonstrated above.

In order to compute the scattering properties of a random rough surface using electromagnetic solvers (Penetrating or Separating) and a Monte Carlo procedure, an ensemble of surface realizations must be generated. There are several ways to generate a statistically stationary random surface [12.48]. The most common approach consists in generating surface profiles by the following technique. A sequence of random numbers ($\sim 10^5$) with Normal statistics, zero mean, and variance (rms roughness) $\sigma = 1$ is constructed from another random series directly generated by a computer. Then the former sequence is scaled in order to obtain a desired σ and,

further, correlation with the Gaussian is performed in order to obtain a profile with a Gaussian correlation function. This is known as the spectral method ([12.38]) and is used in PCGrate codes.

The boundaries of such randomized grating or mirror have both periodical and random roughness components and some averaging of random samples (from a few up to a several hundred) is required to obtain the exact scattered field intensities (see Example 10 in Sec. 12.9). In some conditions, fortunately not in x-ray–EUV, for instance when surface waves can propagate (like polaritons for metallic surfaces in the TM polarization), very big numbers of discritization points and propagating and evanescent orders (about a few thousand or even more) must be taken into account. Rigorous computation of the field scattered by random rough surfaces is a problem of daunting complexity in the area of electromagnetism and optics even for modern computers because of the very small wavelength-to-period and small wavelength-to-height ratios. It is especially true for x-ray–EUV grating and mirror applications. Therefore, the hardest diffraction problems may require large amounts of computer memory and, especially, high speed of computations.

12.8.3 Scattering intensity of rough gratings in a non-dispersive plane

The IMs, which have been developed in the frame of electromagnetic theory, permit application of optical grating methods to analysis of specular and diffuse x-ray–EUV scattering from rough gratings and mirrors using Monte Carlo calculus. The question of the closeness of results for 1D and 2D surfaces is of interest of this Section, since numerical methods for 1D surfaces are well established and efficient, and widely used for surfaces with 2D roughness [12.39]. The derivations of the boundary integral equations using potential operators as well as some details of their numerical implementation were described in previous Sections. An important case of bi-periodic gratings and 2D rough surfaces may be considered in a way by expressing the solution of the 3D Maxwell equations through solutions of the 2D general Helmholtz equations in conical diffraction, an approach which may be resorted to in short waves and shallow surface using the equivalence rule derived in App. D.

The effect of roughness on the 2D DRC can be exactly taken into account with model in which an uneven surface is represented by a bigrating with large periods of $d_{x,z}$ in perpendicular planes, which include appropriate numbers of random asperities with correlation lengths of $\xi_{x,z}$. We analyze a complex structure which, while being the bigrating from a mathematical viewpoint, is actually the rough surface for $d_{x,z} \gg \xi_{x,z}$. If $\xi_{x,z} \sim \lambda$ and the number of orders is large, the continuous angular distribution of the energy reflected from randomly rough boundaries can be described by a discrete distribution η_{mn} in orders (m,n) of a bigrating, similar to (12.134) for classical gratings. A study of the scattering intensity starts with obtaining statistical realizations of profile boundaries of the structure to be analyzed, after which one calculates the DRC for each realization, to end with the DRC averaged out over all realizations to obtain a mean DRC. By selecting large enough samples and numbers of sampling points, one comes eventually to properly averaged properties of the rough surface; however, this approach does not involve approximations, including averaging by the Monte Carlo method.

12.8.3.1 The equivalence rule between 2D and 1D grating efficiencies

A general approach to find efficiencies of bigratings and mean DRCs of rough 2D surfaces which permits one to use exact integral equations, rigorous (extended) boundary conditions,

and radiation conditions leads to tedious calculus even in a case of perfectly conductive surfaces [12.49]. However, a great deal of simplification of the given boundary-problem can be achieved for shallow gratings and randomly-rough surfaces if we use the Rayleigh hypothesis together with the small-amplitude perturbation technique. Implementations of such a method, in which the reduced Rayleigh equations for reflection from such structure are solved in the form of expansions of the amplitudes of the p- and s-polarized components of the scattered field in powers of the surface profile function through terms, up to the third order, were proposed in several papers (see, e.g., Ref. 12.50 and references therein). In the present work, the authors use the perturbative analysis results only in order to derive an approximate connection rule between the efficiency of a shallow bigrating and efficiencies of two classical gratings with grooves rotating on 90 deg. The efficiency itself of a classical grating working in conical diffraction is defined rigorously using the boundary integral equation method, as it is prescribed in previous Sections.

The equivalence rule can be formulated as the following (see (12.163) of App. D)

$$\eta_{mn} = \frac{\eta_m \eta_n}{r_F}, m \vee n = 0, h_{x,z}/d_{x,z} < 1, \tag{12.140}$$

where η_m and η_n are classical grating efficiencies obtained in conical diffraction, $h_{x,z}$—profile heights in perpendicular planes, r_F—the Fresnel factor of a 2D surface. It is worth noting that η_m and η_n in this equivalence rule should be computed with preservation of incidence and polarization angles of both gratings in the absolute coordinate system.

Thus, using (12.140) the efficiency η_{mn} of bigratings can be easily expressed in terms of the product of the efficiencies of two respective classical gratings considered in perpendicular dispersive planes and working in conical mounts at any polarization state. Equation (12.140) was derived in Ref. 12.37 for the normal incidence of linearly-polarized light on a simple boundary-profile bigrating. The equivalence rule described above is very similar to the impulse approximation result of the atomic scattering theory and can include multiple scattering in each perpendicular direction but always excludes cross-correlation components.

The derived connection equation is approximate and valid for shallow periodic surfaces of the type considered. However, this equivalence rule was checked successfully against various numerical examples, including non-shallow bigratings working at different wavelength-to-period ratios [12.37, 12.51]. It was found that it gives accurate results under the following assumptions: (a) $h_{x,z} \lesssim d_{x,z}$ and (b) $\lambda \gtrsim d_{x,z}$. However, for non-deterministic surface profiles working in short waves, some modification of these conclusions is required. As follows from the known results obtained from analytic and asymptotic expressions valid for x-rays (see, e.g., Refs. 12.6, 12.40), (12.140) gives high-accuracy solutions for shallow rough 0D (i.e. rows of atoms with displacements), 1D, and 2D surfaces if the following conditions are fulfilled: (c) $\cos\theta' h_{x,z} \ll d_{x,z}$ and (d) $\lambda \ll d_{x,z}\cos\phi$, where θ' is an incidence angle on the surface. In case of x-ray–EUV ranges refractive indices of materials are close to the vacuum refractive index and $h_{x,z}$ can be large enough for grazing incidence. Thus (a) and (c) are close due to the nature of the perturbative development. However, (d) extends the range of the validation of (12.140) significantly, i.e. to the whole short-wave optical range because of the absence of optical resonances (i.e. due to plasmons, polaritons, waveguide resonances, etc) in x-rays and EUV.

12.9 Examples of numerical results

The described theoretical and numerical approaches for the calculation of far-zone fields and polarization properties of diffraction gratings are well suited to various types of optical grat-

ing analysis. In this Section, we are going to analyze numerically examples of diverse grating diffraction problems. The results presented demonstrate the impact of diffraction and polarization incident angles, boundary shapes and layer refractive indices on diffraction and absorption in periodical structures.

Table 12.1: Diffraction efficiencies (η^+), diffraction (θ^+, ϕ^+) and polarization (δ^+, ψ^+) angles of a metallic lamellar grating[a]

DO[b]	θ^+,°	ϕ^+,°	η^+,%	δ^+,°	ψ^+,°
R_{-2}	−43.715	−20.705	7.52	61.85	48.30
R_{-1}	−9.007	−20.705	13.25	15.79	−12.23
R_0	22.208	−20.705	44.27	41.33	170.15
R_1	65.852	−20.705	31.05	75.64	166.30

[a] $c/d = 0.5$, $2H/d = 1$, $\varepsilon_+ = 1$, $\varepsilon_- = (-24.99, 1)$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 22.208°$, $\phi = 20.705°$, $\delta = 45°$, $\psi = 0$.
[b] Diffraction order

Table 12.2: Diffraction efficiencies ($\eta^\pm$), diffraction $\theta^\pm$, $\phi^\pm$) and polarization($\delta^\pm$, $\psi^\pm$) angles of a dielectric lamellar grating[a]

DO[b]	$\theta^\pm$,°	$\phi^\pm$,°	$\eta^\pm$,%	$\delta^\pm$,°	$\psi^\pm$,°
R_{-2}	35.265	−30	0.1612	64.32	−30.24
R_{-1}	0	−30	0.3807	66.0	−157.22
R_0	35.264	−30	1.854	70.43	−148.60
T_{-3}	−45	−19.471	3.363	51.05	32.28
T_{-2}	−20.705	−19.471	10.35	56.24	110.23
T_{-1}	0	−19.471	31.87	46.54	99.02
T_0	20.705	−19.471	14.19	34.26	68.38
T_1	45	−19.471	37.83	46.34	86.83

[a] $c/d = 0.5$, $2H/d = 0.5$, $\varepsilon_+ = 1$, $\varepsilon_- = 2.25$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 35.264°$, $\phi = 30°$, $\delta = 45°$, $\psi = 90°$.
[b] Diffraction order

In this Section, we present several numerical experiments taken from well-known spectroscopic applications of gratings working in various mounts and polarization states at different wavelengths. More specifically, they are: the typical dielectric and metallic lamellar gratings illuminated in conical diffraction; the typical dielectric sine grating working in off-plane mounts; the typical metallic echelette gratings illuminated in conical diffraction; the anomalously absorbing Ag shallow-sine grating working in off-plane mounts in the visible; the photonic crystals with Au nanorods of various cross-sections illuminated at normal incidence in the visible–near-infrared; the photonic crystal with dielectric circular nanorods working in different mounts in the near- and mid-infrared; the Al echelle grating protected by a thin layer of MgF_2 and illuminated in conical diffraction in the vacuum ultraviolet (VUV); the Au blaze grating working in grazing-incidence off-plane mounts in soft x-rays; the minimally-absorbing Mo/B_4C multilayer blaze grating illuminated in grazing conical diffraction in soft x-rays; the flight Mo/Si multilayer trapezoidal grating working in the near-normal-incidence EUV and with random roughness accounting. The numerical examples of calculation results described in this Section were calculated using a few commercial and non-commercial IM-based codes.

12.9.1 Efficiencies and polarization angles of lamellar gratings

The efficiency results of reflection orders of the present IM for a typical conducting lamellar grating with the ridge width c and depth $2H$ working in a conical mount are demonstrated in Table 12.1. The grating and light parameters are as follows: $c/d = 0.5$, $2H/d = 1$, $\varepsilon_+ = 1$, $\varepsilon_- = (-24.99, 1)$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 22.208°$, $\phi = 20.705°$, $\delta = 45°$, and $\psi = 0$. We used 400 discretization points, mesh grading and the discretization of V^+J^- to calculate this example that allocates 188 MByte memory. The energy balance error calculated from (12.74) is $\sim 10^{-6}$. The average time taken up by the example on a workstation with two Quad-Core Intel® Xeon®2.66 GHz processors, 8 MB L2 Cache, 1333 MHz FSB and 16 GB RAM is ~ 1.5 s when operating on Linux Ubuntu 12.04 LTS 64-bit or Windows Vista® Ultimate 64-bit and employing eightfold paralleling.

Table 12.3: Diffraction efficiencies ($\eta^\pm$) and diffraction ($\theta^\pm$, $\phi^\pm$) and polarization ($\delta^\pm$, $\psi^\pm$) angles of a dielectric sine grating for $B_z = 0$[a]

DO[b]	$\theta^\pm$,°	$\phi^\pm$,°	$\eta^\pm$,%	$\delta^\pm$,°	$\psi^\pm$,°
R_{-3}	−43.384	−15	1.121	70.99	3.60
R_{-2}	−9.744	−15	3.741	26.90	0.93
R_{-1}	20.389	−15	3.873	63.25	178.18
R_0	60	−15	10.33	88.93	178.05
T_{-5}	−57.013	−7.435	.01855	80.19	−114.68
T_{-4}	−35.921	−7.435	.002482	52.58	100.24
T_{-3}	−19.545	−7.435	.7394	57.61	−179.28
T_{-2}	−4.729	−7.435	4.922	22.90	174.84
T_{-1}	9.770	−7.435	9.923	60.39	4.72
T_0	24.949	−7.435	7.145	77.32	6.84
T_1	42.371	−7.435	51.83	84.43	−5.78
T_2	67.826	−7.435	6.351	84.85	−11.39

[a] $2H/d = 0.3$, $\varepsilon_+ = 1$, $\varepsilon_- = 4$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 60°$, $\phi = 15°$, $\delta = 81.501°$, $\psi = 0$.
[b] Diffraction order

In Table 12.2, the efficiency data of reflection and transmission orders for a similar dielectric lamellar grating in a conical mount are presented. The grating and light parameters are as follows: $c/d = 0.5$, $2H/d = 0.5$, $\varepsilon_+ = 1$, $\varepsilon_- = 2.25$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 35.264°$, $\phi = 30°$, $\delta = 45°$, and $\psi = 90°$. We used $N = 400$, mesh grading and the discretization of V^+J^- to calculate this example that allocates 188 MByte memory. The energy balance error calculated from (12.74) is $\sim 10^{-5}$. The average time taken up by the example is ~ 1.5 s when operating on the aforementioned workstation and operating system. The efficiencies and polarization angles obtained in this and two next Subsections for transmission and reflection gratings working in conical diffraction can be compared with those obtained by the use of other rigorous methods and codes [12.7, 12.8].

12.9.2 Efficiencies and polarization angles of dielectric sine grating

In Tables 12.3 and 12.4, the efficiency results of the IM for a typical dielectric sine grating working in a conical mount are presented. The grating and light parameters are as follows:

Table 12.4: Diffraction efficiencies ($\eta^{\pm}$) and diffraction ($\theta^{\pm}$, $\phi^{\pm}$) and polarization ($\delta^{\pm}$, $\psi^{\pm}$) angles of a dielectric sine grating for $E_z = 0$[a]

DO[b]	$\theta^{\pm}$,°	$\phi^{\pm}$,°	$\eta^{\pm}$,%	$\delta^{\pm}$,°	$\psi^{\pm}$,°
R_{-3}	−43.384	−15	1.121	70.99	3.60
R_{-2}	−9.744	−15	3.741	26.90	0.93
R_{-1}	20.389	−15	3.873	63.25	178.18
R_0	60	−15	10.33	88.93	178.05
T_{-5}	−57.013	−7.435	.01855	80.19	−114.68
T_{-4}	−35.921	−7.435	.002482	52.58	100.24
T_{-3}	−19.545	−7.435	.7394	57.61	−179.28
T_{-2}	−4.729	−7.435	4.922	22.90	174.84
T_{-1}	9.770	−7.435	9.923	60.39	4.72
T_0	24.949	−7.435	7.145	77.32	6.84
T_1	42.371	−7.435	51.83	84.43	−5.78
T_2	67.826	−7.435	6.351	84.85	−11.39

[a] $2H/d = 0.3$, $\varepsilon_+ = 1$, $\varepsilon_- = 4$, $\mu_{\pm} = 1$, $\lambda/d = 0.5$, $\theta = 60°$, $\phi = 15°$, $\delta = 8.499°$, $\psi = 180°$.
[b] Diffraction order

$2H/d = 0.3$, $\varepsilon_+ = 1$, $\varepsilon_- = 4$, $\mu_{\pm} = 1$, $\lambda/d = 0.5$, $\theta = 60°$, $\phi = 15°$. For Table 12.3, the incident polarization angles are $\delta = 81.501°$ and $\psi = 0$, for Table 12.4—$\delta = 8.499°$, $\psi = 180°$.

We used 100 discretization points and the numerical differentiation of V^+ to calculate these examples which allocate 10 MByte of RAM. The energy balance error calculated from (12.74) is about 10^{-5} for both components of the polarization incident radiation. The average computation time taken up by an example on the aforementioned workstation and operating system is ~ 0.1 s.

12.9.3 Efficiencies and polarization angles of metallic echelette grating

The numerical results for a typical metallic echelette grating with blaze angle ζ and an apex angle of 90° (see Fig. 12.2) working in a conical mount are demonstrated in Tables 12.5 and 12.6 for the two basic states of the incident polarization: $\delta = 0$, $\psi = 180°$ or $\delta = 90°$, $\psi = 0$. The grating and light parameters are as follows: $\zeta = 30°$, $\varepsilon_+ = 1$, $\varepsilon_- = (-45, 28)$, $\mu_{\pm} = 1$, $\lambda/d = 0.5$, $\theta = 0$, $\phi = 40°$, and $\psi = 0$. One has used $N = 800$, mesh scaling near edges and the differentiation of V^+ to calculate these examples allocating 196 MByte of RAM. The average energy balance error calculated from (12.74) is $\sim 10^{-5}$ for both polarization states of the incident radiation. The average computation time taken up by two values of the polarization angle on the aforementioned workstation and operating system is ~ 3 s.

12.9.4 Anomalously absorbing Ag shallow-sine grating in the visible

Resonance and non-resonance anomalies differing in their nature can be effectively explored in high conductive gratings, such as: surface plasmon excitations, Bragg and Brewster conditions, groove shape features, etc. Because the *s* and *p* modes in conical diffraction are coupled through the boundary conditions, the associated problems are more general, and gratings act as perfect absorbers and local-field enhancers.

Table 12.5: Diffraction efficiencies (η^+) and diffraction (θ^+, ϕ^+) and polarization (δ^+, ψ^+) angles of a metallic echelette grating for $\delta = 0$, $\psi = 180°$[a]

DO[b]	θ^+,°	ϕ^+,°	η^+,%	δ^+,°	ψ^+,°
R_{-1}	−40.746	−40	12.97	39.447	−175.93
R_0	0	−40	28.49	86.414	−50.97
R_1	40.746	−40	24.81	39.209	7.67

[a] $\zeta = 30°$, $\varepsilon_+ = 1$, $\varepsilon_- = (-45, 28)$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 0$, $\phi = 40°$.
[b]Diffraction order

Table 12.6: Diffraction efficiencies (η^+) and diffraction (θ^+, ϕ^+) and polarization (δ^+, ψ^+) angles of a metallic echelette grating for $\delta = 90°$, $\psi = 0$[a]

DO[b]	θ^+,°	ϕ^+,°	η^+,%	δ^+,°	ψ^+,°
R_{-1}	−40.746	−40	53.15	54.0	13.37
R_0	0	−40	17.48	4.58	95.21
R_1	40.746	−40	9.444	49.41	−171.22

[a] $\zeta = 30°$, $\varepsilon_+ = 1$, $\varepsilon_- = (-45, 28)$, $\mu_\pm = 1$, $\lambda/d = 0.5$, $\theta = 0$, $\phi = 40°$.
[b]Diffraction order

In Fig. 12.17, the absorption of the Ag sinusoidal grating with $d = 2.2$ μm and $2H =$ 100 nm is calculated for the $\delta = 90°, \psi = 0$ or $\delta = 0, \psi = 180°$ polarized incidence light with $\lambda = 663$ nm as a function of θ for $\phi = 0$ (classical, TE and TM) or $\phi = 50°$ (conical). The refractive indices of Ag were taken from Ref. 12.36 ($\mu_\pm = 1$). For in-plane diffraction, anomalous absorption exists only for the TM polarization, while for conical diffraction both components are absorbed but in smaller amounts.

Note that we used the variant of discretization of H^+V^- to calculate these examples. The calculated problem allocates 10 MByte of RAM using $N = 100$. The energy balance error calculated from (12.74) is about 10^{-6} for both components of the polarization of incident radiation. The average computation time taken up by the example on the aforementioned workstation and operating system is less than 0.1 s per calculation point.

12.9.5 Photonic crystals with Au nanorods in the visible–near-IR

In this Subsection, we are going to analyze numerically the optical response (reflection and absorbtion) of photonic crystal slabs supporting polariton-plasmon propagation with different cross sections of nanowires invariant with respect to the z axis. The essential physics of the formation of localized plasmon polariton modes in metallic nanowire arrays is described in Chapter 1. The vital role of the absorption, slab cross-section shape, and filling ratio of photonic crystals in the visible and near infrared regions is demonstrated in this Subsection. The model contains $M-2$ (see Fig. 12.3) identical gratings with closed boundaries (inclusions) of simple cross sections displaced vertically (by h_m) and horizontally (by f_m) relative to one another and embedded in a homogeneous medium with dielectric permittivity ε_1 and magnetic susceptibility μ_1. We deal here only with materials with $\mu_m = 1$, $m = 0, \ldots M$, although the model is applicable to other cases as well, including metamaterials [12.18]. The dependence of the dielectric permittivity ε_m, $m = 2, \ldots M-1$ of the material of nanorods on the incident

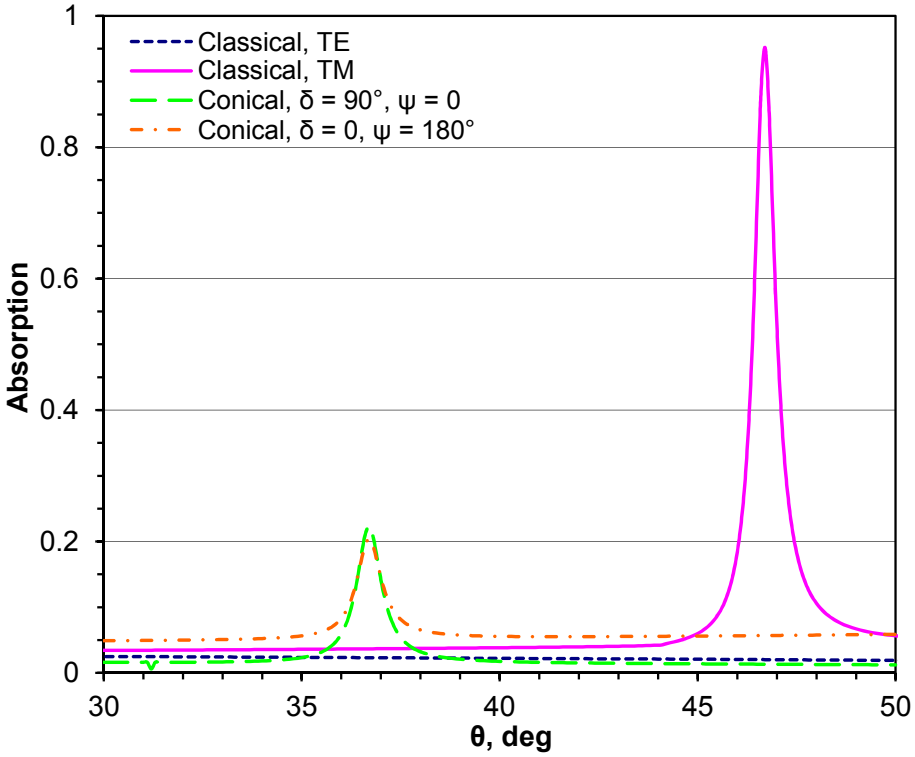


Figure 12.17: Absorption of an Ag sinusoidal grating with $d = 2.2\mu m$ and a depth of 100 nm working in classical ($\phi = 0$) or conical ($\phi = 50^\circ$) diffraction, plotted vs. θ for $\lambda = 663$ nm and $\delta = 90^\circ, \psi = 0$ or $\delta = 0, \psi = 180^\circ$.

photon frequency is assumed to be known. The lower medium and the upper medium are likewise assigned pairs of material constants, but one may conceive of more complicated cases of multilayer structures as well. The model also allows arbitrary incidence of, in the general case, elliptically polarized radiation on photonic crystals, which is prescribed by two angles of incidence and two angles of polarization.

Figure 12.18 displays for comparison theoretical spectra of energy reflected from, and absorbed by, a photonic crystal with Au nanowires of circular, square, rectangular, and triangular cross sections of the same area and with $M = 3$ studied in the $1-3$-eV photon energy range (visible and near infrared). In this and similar subsequent examples, we consider the TM-polarized ($\theta = \phi = \delta = 0$, $\psi = 180^\circ$) light normally falling on Au nanowires embedded in a SiO_2 matrix with $d = 200$ nm, $\varepsilon_0 = \varepsilon_1 = \varepsilon_3 = 2.13$, and refractive indices of Au taken from Ref. 12.36. The orientation of the rods having edges is chosen in such a way that light normally falls on one side of the rods. The $a \times b$ dimensions of the rectangular rods selected for this example are 50×25 nm^2 or 25×50 nm^2 and the width of the squares or triangles and diameter of the circles were chosen to obtain equal cross sectional area $S = 1250$ nm^2. As seen from Fig. 12.18, reflection and, particularly, absorption spectra exhibit a strong difference near the plasmon-polariton anomaly among the five shapes of the nanowire cross section chosen. These differences amount to several hundred percent for the rectangles because of their different width-to-height ratio (two and a half) compared with the square or the circle (one) and the equilateral triangle (0.866). One observes also a noticeable difference in the positions of the absorption and reflection maxima among different grating profiles. Thus, the simple effective medium theory cannot be applied to design and analysis of such photonic crystals, even for a small filling ratio [12.13].

Figure 12.19 presents energy spectra similar to those displayed in Fig. 12.18 but for S four times that of the preceding example. In this case, $a \times b = 100 \times 50$ nm^2 or 50×100 nm^2. We readily see that the differences in the reflection and absorption spectra among gratings of different profiles increase with increasing filling ratio and are observed now not only close to the plasmon resonances. Near the resonances, they amount to a few dozen percent of energy. The absorption spectra of the triangular-shaped nanowires have an interesting step-like function behaviour, which is not the case for absorption spectra of nanowires of other rod shapes.

Only 50 discretization points, mesh grading, Hankel kernel functions for inclusions and discretization of H^+V^- have been used to compute these examples which allocate ~ 0.1 MByte memory. The relative error calculated from the energy balance for absorption gratings is $\sim 10^{-4}$. The average time taken up by one point on the aforementioned workstation and operating system is less than 0.1 s.

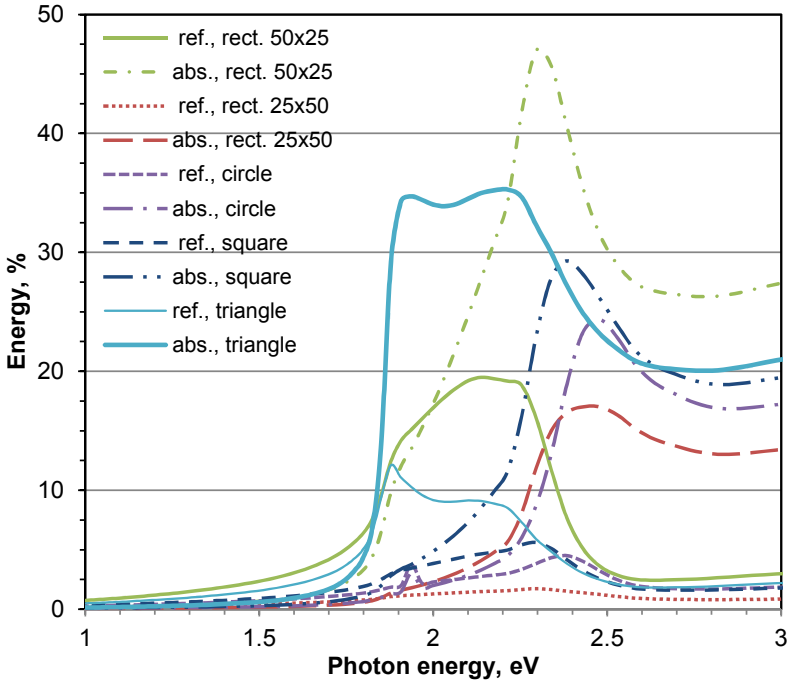


Figure 12.18: Calculated reflection (ref.) and absorption (abs.) spectra of SiO_2-embedded gratings with $d = 200$ nm and a layer of different Au-nanowire cross sections of the same area of $S = 1250$ nm^2, plotted vs. photon energy for normal incidence and TM polarization.

12.9.6 Lossless photonic crystal with circular rods in the near- and mid-IR

In this example, we consider numerically some diffraction properties of non-absorbing photonic crystals with dielectric rods. The influence of the geometry and number of crystal layers, the shape of rods, the filling ratio, the index of refraction of materials and the polarization and diffraction angles of light can be investigated for this type of photonic crystals. The role of the filling ratio, refractive index and polarization was demonstrated for the classical diffraction [12.12, 12.31]. Here we demonstrate, as an example of possibilities of developed software, the vital role of the filling ratio and polarization for conical diffraction.

Figures 12.20 and 12.21 display spectral transmission for photonic crystal circular rods with $d = 1\,\mu$m and $\varepsilon_m = \varepsilon_{rod} = 4$, $m = 2 \ldots M-1$, $\mu_m = 1$, $m = 0, \ldots M$ embedded in vacuum

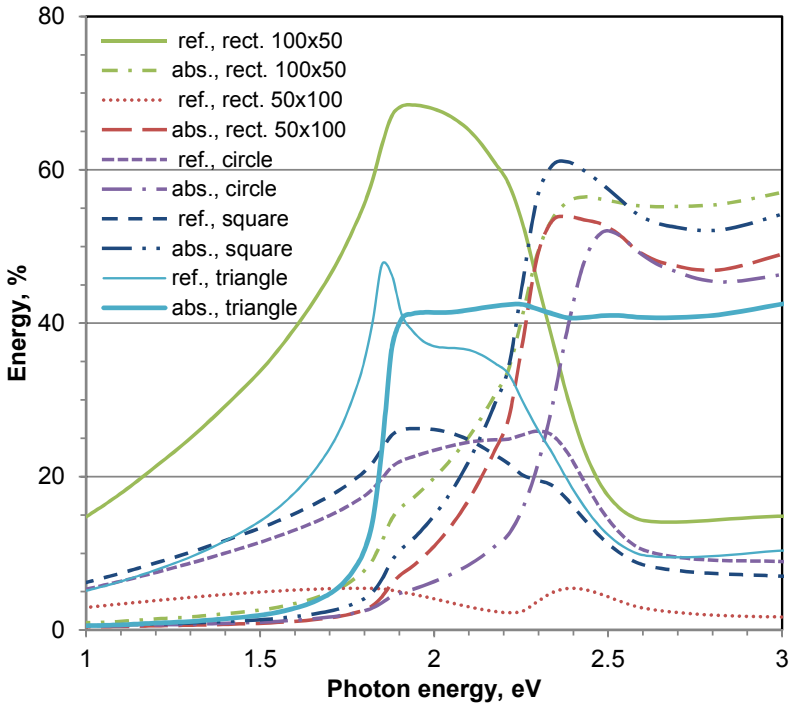


Figure 12.19: The same as in Figure 12.18, but for $S = 5000\ nm^2$.

($\varepsilon_0 = \varepsilon_1 = \varepsilon_M = 1$) at filling ratios of $p = 0.125$ and $p = 0.5$ for $M = 17$, $h_m = 0.866\,\mu$m, and $f_m = 0.5\,\mu$m (hexagonal crystal geometry) for $\theta = 0$ and $\delta = 90°$, $\psi = 0$ or $\delta = 0°$, $\psi = 180°$ (see the detailed model description in the previous numerical example). In Fig. 12.20 one can see in-plane diffraction efficiencies ($\phi = 0$) and similar transmittance data were computed in Ref. 12.31 by the boundary integral equation method of Ch. 4 (Figs. 6 and 11 of Ref. 12.31). In Fig. 12.21 for the off-plane diffraction $\phi = 30°$ and this is an additional parameter compared with the classical diffraction case.

For both in-plane and off-plane examples, there is a very different behavior in diffraction properties for TE and TM polarization components of the incident radiation, especially for big filling ratios. Compared with respective curves obtained in Figs. 12.20 and 12.21, it emerges that for s-polarized light the centers of the conical diffraction gaps have shifted significantly to smaller wavelengths and the widths and depths of the gaps have decreased considerably. In contrast to this behavior, for p-polarized light the centers of the conical diffraction gaps compared with the in-plane ones have shifted a little bit in opposite directions and the widths and depths of these gaps have increased noticeably. The vital importance of the azimuthal angle ϕ, as well as the incidence polarization has become evident even for a small filling ratio ($p = 0.125$); however they are more important for a high filling ratio ($p = 0.5$). Thus, using the conical diffraction for dielectric photonic crystals gives additional control parameters which significantly affect Bragg diffraction and existing photonic band gaps.

Only $N = 50$ without mesh grading and with Hankel kernel functions for inclusions are required to compute these examples using discretization of H^+V^- which allocates ~ 0.2 MB memory. The relative error calculated from the energy balance for non-absorption gratings is $\sim 10^{-4}$. The average time taken up by one point on the aforementioned workstation and operating system is less than ~ 0.1 s.

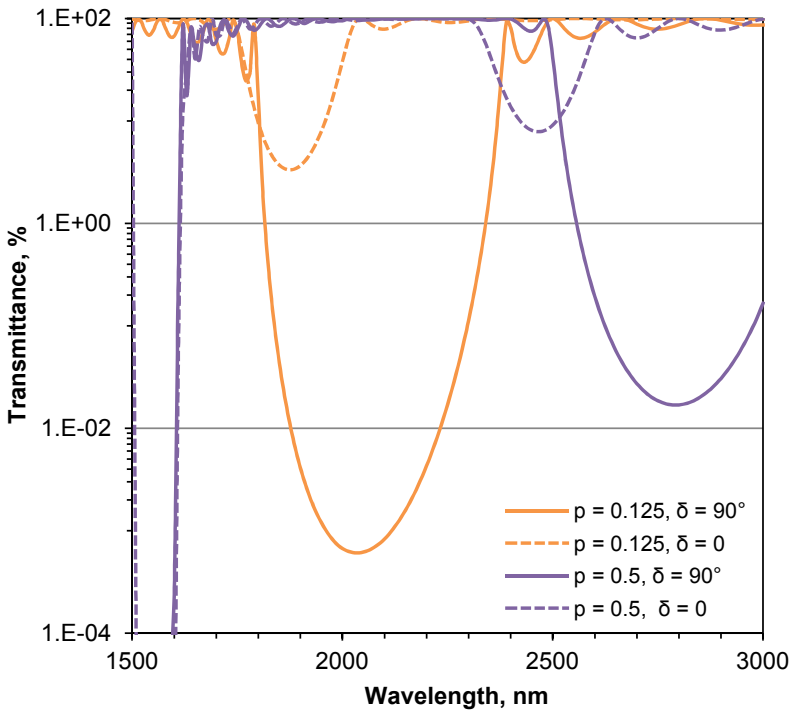


Figure 12.20: Calculated transmission spectra of $1\,\mu$*m-period gratings with 15 layers of dielectric circular rods with* $\varepsilon = 4$ *and different filling ratios p embedded in vacuum with hexagonal structure, plotted vs.* λ *for* $\theta = 0$*,* $\phi = 0$ *and different polarization angles (classical diffraction).*

12.9.7 Al echelle grating coated by MgF_2 in the VUV

Echelle gratings or simple echelles working in high spectral orders near Littrow diffraction conditions at high angles θ are one of the most popular grating types; however, they are rather difficult for fabrication and efficiency computations, especially those with dielectric coatings. A thin oxide film on the Al grating surface may lead to degradation of its diffraction properties at wavelengths below 130–140 nm. To protect and even improve the echelles' reflectance surfaces, a thin dielectric coating with a thicknesses of a few dozen nm can be applied in the VUV range. The usual material is MgF_2, but sometimes other dielectrics are used. At a certain thickness of the coating film, waveguide phenomena come out to affect the grating performance; as a result, the diffraction efficiency can either decrease or increase as compared to the non-oxidized bare grating. A non-conformal layer which is obtained by two adjacent non-parallel boundaries (having different vertical distances between) provides a new freedom in design, but the analysis of gratings becomes more complex. Furthermore, echelles are frequently used in conical diffraction, making it possible to separate beams in a non-dispersive plane [12.3].

Our example deals with an aluminium echelle with 316 grooves/mm, working blaze angle $\zeta_1 = 63.4^\circ$ (r-2, i.e. $\tan\zeta_1 = 2$), and apex angle 90°. The grating works at the -47th order, wavelength $\lambda = 120$ nm and $\phi = 6.5^\circ$. A protecting MgF_2 layer is applied. Other coating and light parameters are as follows: $\varepsilon_0 = 1$, $\varepsilon_1 = \varepsilon_{MgF_2} = (2.643876, 0)$, $\varepsilon_2 = \varepsilon_{Al} = (-1.2353087, 0.0913816)$, $\mu_m = 1$, $m = 0, \ldots 2$, $\delta = 0$, and $\psi = 180^\circ$. We consider four variants of its thickness and shape including zero thickness for the bare Al grating. For coated gratings, the coating's upper boundary is sawtooth, with right angle at the top vertex situated $h_0 = 30$ nm above the grating's top vertex. Thus, the variants differ from each other by the coating's working angle, which is $\zeta_0 = \zeta = 0$ for the bare case, $\zeta = 63.4^\circ$—for the conformal

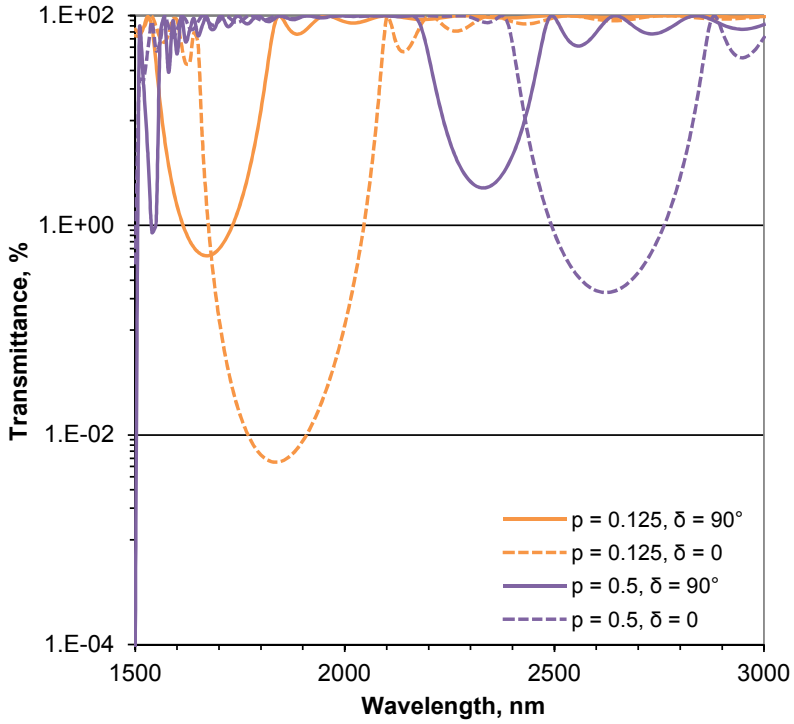


Figure 12.21: The same as in Figure 12.20, but for $\phi = 30°$ (conical diffraction).

case and $\zeta = 62.9°$ or $\zeta = 63.9°$—for two non-conformal cases. Such non-conformal models of coatings do not pretend to be the best description of real structures formed by sputtering, but are simple and possible; and they account for a deviation of coating direction from the Al substrate surface which leads to a tapered shape on both slopes of the triangular profile.

Fig. 12.22 presents angular dependencies of the grating efficiency. The efficiency results for such echelles obtained using different IM-based codes in in-plane mounts are presented in Ref. 12.25. Fig. 12.22 shows that the conformal coating leads to a noticeable increment of efficiency in comparison with a bare case over the whole range of angles, by $\sim 20\%$. The non-conformal coating with $\zeta = 62.9°$ increases the efficiencies by $\sim 10\%$ compared to the bare grating. The geometry in this case is such that the working facet receives a thinner layer of MgF_2, which narrows approaching the vertex; the non-working facet gets a fatty coating. In contrast, the non-conformal coating with working angle $\zeta = 63.9°$ does not increase the efficiency at its maximum and leads to practically the same efficiency graph as for the bare Al grating case in the whole central angular range. The opposite impact of these non-conformal coatings working in classical diffraction is demonstrated in Ref. 12.25. Thus, the efficiency is very sensitive to the boundary vertical shift, to the deviation of a MgF_2 layer from conformal shape, and also to the off-plane deviation.

Computations in this example were carried out with $N = 800$ for the bare grating and with $N = 1600$ for the gratings with conformal and non-conformal coatings. One also has used mesh scaling near edges and the differentiation of V^+ to calculate these examples, allocating 1024 MByte of RAM for $N = 1600$. The relative error calculated by (12.128) from the energy balance for absorption gratings is $\sim 10^{-4}$. In case of piecewise linear profiles, many pairs of kernel function arguments can be obtained from each other by translations; corresponding kernel function values are equal. Hence, there is an effective way to check for given arguments, whether or not we already encountered a congruent pair and calculated the kernel function for

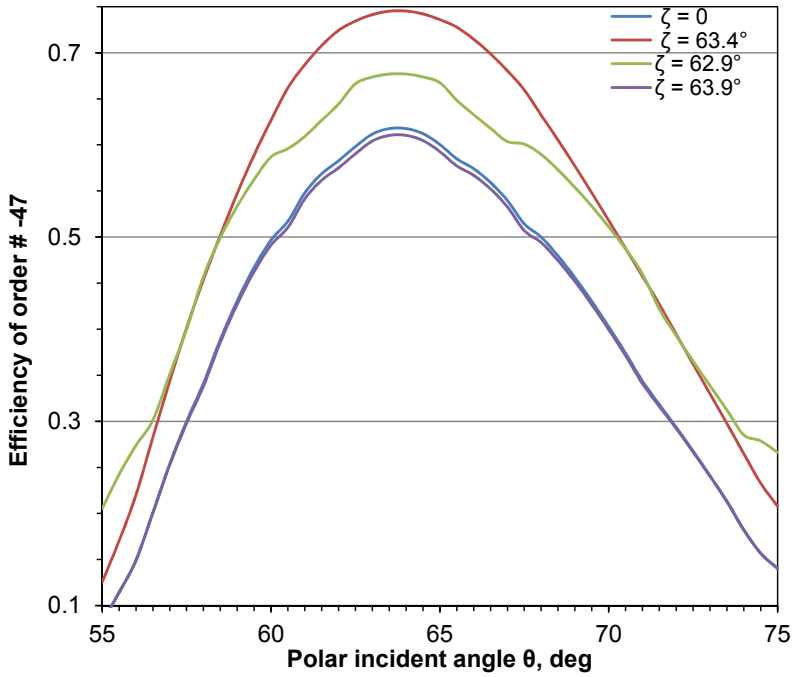


Figure 12.22: Efficiency in −47 order of A1 316 grooves/mm echelles with blaze angle $\zeta_1 = 63.4°$ working in conical diffraction at $\lambda = 120$ nm and $\phi = 8°$: bare ($\zeta = 0$), or with MgF_2 coating having upper sawtooth boundary with vertical displacement $h_0 = 30$ nm and working angle $\zeta = 63.4°$ (conformal case), or $\zeta = 62.9°$ (non-conformal case), or $\zeta = 63.9°$ (non-conformal case), plotted vs. θ.

it (see Sec. 12.6.2). This approach significantly reduces computational time for echelles and even more—in case of conformal layers, where the kernel function values calculated on the upper side of the layer can be reused on the lower side. Calculation for each point on Fig. 12.22 required between a few s (bare case) and several dozen s (conformal and non-conformal cases) on the aforementioned workstation and operating system.

12.9.8 Au off-plane-grazing-incidence blaze grating in soft x-rays

The conical diffraction mount in which the direction of incident light is confined to a plane parallel to the direction of the grooves has the unique property of maintaining high and sustained diffraction efficiency, which is very important in the x-ray–EUV range. Such gratings are utilized as dispersive elements in laboratory and space spectral instruments, time-delayed compensators or splitters and spectral purity filters for EUV lithography. Grazing-incidence off-plane gratings have been suggested for the International X-ray Observatory (IXO) [12.53]. Compared with gratings in the classical in-plane mount, x-ray gratings in the off-plane mount have the potential for superior resolution and efficiency for the IXO mission. The results of efficiency calculations for such a gold blazed soft x-ray grating in a conical mount using the perfect triangular groove profile with $d = 200$ nm are shown in Fig. 12.23. The design blaze angle ζ is 7.5° and the technique anti-blaze angle is 64.53° [12.54]. Remaining grating and light parameters are as follows: $\mu_\pm = 1$, $\theta = 0$, $\phi = 88°$, and $\delta = 90°$ and $\psi = 0$ or $\delta = 0$ and $\psi = 180°$.

In Fig. 12.23, the numerical results of the IM presented for a finite boundary conductivity are compared with those based on the IM with the perfect boundary conductivity multiplied by Fresnel reflectances calculated with respect to the blaze facet. The incident beam in the computations based on the perfect conductivity model and classical diffraction (using the In-

variance theorem (see in [12.37] and Ch. 4) was assumed to be 100% TE-polarized ($B_z = 0$). The refractive indices of Au were derived from the compilation at [12.55].

Rigorous computations carried out by the methods presented show that for the finite grating model all the order efficiencies are not sensitive to a polarization state. For both basic polarization state of the incident radiation order efficiencies presented in Fig. 12.23 differ not more than a few tenths of a %. Contrary, calculations based on the perfectly conducting boundary model are very sensitive to the polarization state and sharp Rayleigh anomalies for the TM-polarized incident radiation (not shown) occur. As can be seen in Fig. 12.23, the agreement between the data obtained by the finite conductivity model and the perfect conductivity model is good when the TE-polarization is used for the perfect conductivity model. The same conclusions were derived for a similar grating problem in Ref. 12.7 using the real (measured) average groove profile for the efficiency computation.

We have used 800 discretization points, the numerical differentiation of V^+ and no mesh scaling to calculate the finite-conducting blaze-groove-profile example that allocates a space of 144 MByte. The energy balance error calculated from (12.74) is $\sim 10^{-4}$ in the investigated wavelength range. The average computation time taken up by one wavelength on the aforementioned workstation and operating system is ~ 2 s. The time of a computation using the perfect conductivity model for $N = 200$ is about eighty times shorter at the same computation accuracy.

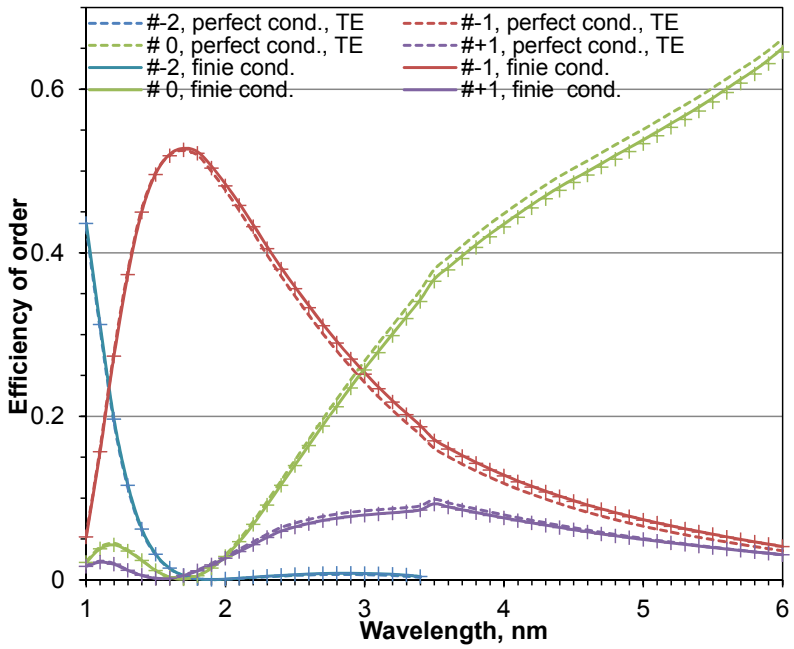


Figure 12.23: Diffraction efficiencies of an Au triangular-groove-profile grating with $d = 200$ nm, $\zeta = 7.5°$, $\mu_\pm = 1$ and for the incident wave with $\theta = 0$, $\phi = 88°$ and $\delta = 90°$, $\psi = 0$ or $\delta = 0$, $\psi = 180°$, plotted vs. λ.

12.9.9 W/B_4C multilayer off-plane-grazing-incidence blaze grating in soft-x-rays

Multilayer coated blazed gratings with high groove density are the best candidates for use in high resolution EUV and soft x-ray spectrometry such as resonance inelastic x-ray spectroscopy. Theoretical and experimental analysis show that such a grating can be potentially optimized for

high dispersion and spectral resolution in a desired high number diffraction order without significant loss of diffraction efficiency. In order to realize this potential, the grating should have a perfect triangular groove profile and its absorption should be minimized. The grazing-incidence conical-diffraction mounting in which the direction of incident light is confined to a plane parallel to the direction of the grooves has the unique property of maintaining a maximal level of diffraction efficiency due to an additional angular parameter. In this Subsection, we analyze the optical absorption of a blazed multilayer grating working in grazing conical diffraction in the soft x-ray range.

In Fig. 12.24, the absorption of the 10000 /mm blazed Si grating coated with 60 bi-layers of W/B_4C is calculated for the polarized ($\delta = 90°$, $\psi = 0$) incidence radiation with $\lambda = 1.3$ nm and $\theta = 6°$ as a function of the azimuthal angle ϕ. The grating has a triangular groove profile with the blaze angle of 6° and antiblaze angle of 64.53° and a conformal multilayer coating (see Sec. 12.9.7) with the thicknesses of W and B_4C layers measured in respect to the working facet normal, 0.6006 nm and 2.4024 nm , respectively. The refractive indices of Si, W, and B_4C were taken from [12.55]. Figure 12.24 displays for comparison theoretical absorption spectra of a Si mirror coated with the same multilayer and working in the same mount. As one can see in Fig. 12.24, for the defined polar angle the grating and mirror absorptions are close in the azimuthal angle range investigated. Grating absorption minima less than 70% can be obtained for the azimuthal angle of $\sim 77.2°$. Thus, almost the all reflected energy can be directed into diffraction orders without additional losses for the multilayer soft-x-ray grating absorption.

Only $N = 400$ was used to compute this grating example accounting 121 boundaries which allocates ~ 60 MB of RAM. The relative error calculated from the energy balance using Eq. (12.128) is $\sim 10^{-4}$. The average time taken up by one point on the aforementioned workstation and operating system is ~ 4 min.

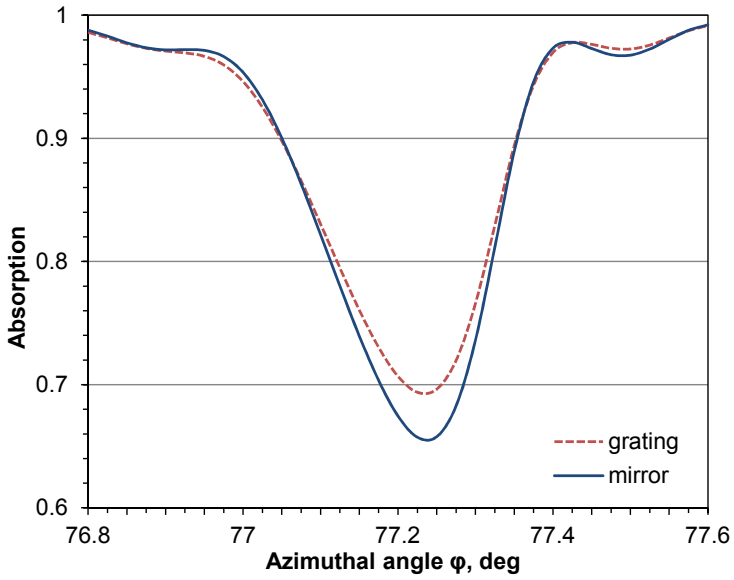


Figure 12.24: Absorption of structures with 60 W/B_4C bilayers on Si for the polarized ($\delta = 90°$, $\psi = 0$) grazing incidence x-ray radiation with $\lambda = 1.3$ nm and $\theta = 6°$ vs. ϕ.
1—mirror; 2—- blazed grating with 10000 grooves/mm and $\zeta = 6°$.

12.9.10 Flight Mo/Si multilayer rough lamellar grating in the EUV

Here we present examples of the Mo/Si lamellar grating efficiency standardized for the Extreme-Ultraviolet Imaging Spectrometer (EIS) on the Hinode (former Solar-B) mission [12.56], the first implementation of a multilayer grating on a satellite instrument. We describe the performance of the flight FL1 4200 grooves/mm multilayer grating operating at $\theta = 6.5^\circ$ of the in-plane configuration in the wavelength region 17–21 nm. The efficiency was calculated by PCGrate-SX v.6.5 software using data of AFM measurements and was compared to the synchrotron efficiency measurements [12.1].

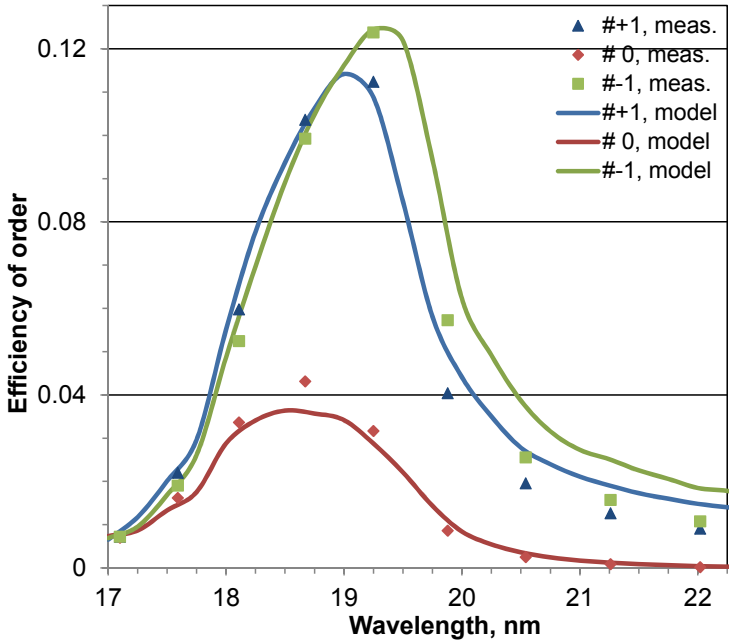


Figure 12.25: Calculated TM efficiencies of orders of a 4200 *grooves/mm rough trapezoidal grating with 20 Mo/Si bilayers on Si operating at* $\theta = 6.5^\circ$ *vs.* λ.

The depth of all the boundary profiles of the multilayer grating was 6.0 nm, with side slopes of 35° and equal top and groove widths, as derived from the AFM and efficiency measurements. Because polarization effects are small near normal incidence, the efficiencies are presented for the case of TM-polarized radiation ($\phi = \delta = 0$, $\psi = 180^\circ$). To determine the absolute values of order efficiencies, a model of the two-period-randomized-trapezium grating describing the realistic boundary shape and roughness was applied. For a rigorous accounting of the random roughness impact on the efficiency, the model with 41 randomly rough borders of the period of ~ 476.19 nm having 400 random sampling points on two trapezoidal grooves with the same Gaussian surface roughness height statistics and Gaussian autocorrelation function was applied (for random border generation on non-flat surface shapes, see [12.1]) . The rough boundary parameters are as follows: the Si-Mo interface rms roughness $\sigma_{\mathrm{Si-Mo}} = 0.2$ nm and the Mo-Si rms roughness $\sigma_{\mathrm{Mo-Si}} = 0.85$ nm. The lateral correlation length $\xi = 5$ nm was chosen from the detailed microscopic analysis and the growth model of typical Mo/Si layers obtained by using magnetron sputtering [12.47]. An assumption about the absence of a vertical correlation between the border random roughness components was applied in this model. Seven sets of 41 rough border profiles were generated to compute exact efficiencies of the FL1 multilayer grating. The Si protective capping layer of 2 nm was modeled by using 1.5-nm-thick

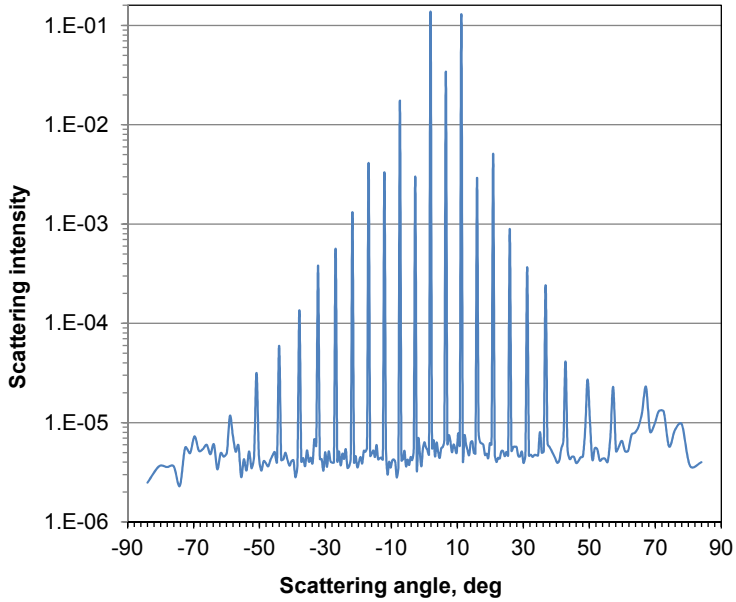


Figure 12.26: Calculated TM scattering intensities of a 4200 *grooves/mm rough trapezoidal grating with 20 Mo/Si bilayers on Si operating at* $\theta = 6.5^\circ$ *vs.* λ.

amorphous SiO_2 on 1.5-nm Si in order to account for the oxidation of the Si capping layer. The FL1 multilayer parameters extracted from the mirror investigation are as follows: 20 Mo/Si layer pairs with the bilayer period $D = 10.3$ nm, Mo thickness to D ratio $\Gamma = 0.37$.

To determine the absolute values of scattering light intensities between orders [12.5], a model of ten-period-randomized-trapezium grating allowing a fine-diffraction-angle discretization and describing the realistic boundary shape and roughness was applied. For a rigorous accounting of random roughnesses, the model with 41 randomly rough low-frequency borders of the period of ~ 2381 nm having 800 random sampling points on 10 trapezoidal grooves with the described above rough boundary parameters was used. Some 105 sets of 41 non-correlated vertically border profiles were generated to compute exact scattering light intensities between orders of the FL1 multilayer grating. The same layer parameters as for the efficiency model (see above) were used. Refractive indices derived from the NIST data for Mo [12.57], from the CXRO data—for Si [12.53], and from the Palik data—for SiO_2 were used for efficiency calculations in the whole wavelength range.

Convergence and accuracy of the efficiency results of the randomly-rough 41-boundary grating were investigated using the Penetrating solver, Gauss computation algorithm and finite type of low border conductivity. All the accelerating convergence options were switched on in PCGrate-SX v. 6.5 (see Sec. 12.4.3). The linear type of refractive index data interpolation was chosen. A high rate of convergence of the results was observed for the developed grating efficiency model [12.1]. Only several sets of 41 rough border profiles and the medium number of discretization points per boundary are enough to compute exact efficiencies in all orders of interest. The differences between principal order efficiencies obtained with seven boundary sets with low ($N = 600$), medium ($N = 800$), and high ($N = 1000$) accuracy are about a few percents for all orders under study. The differences between efficiencies obtained with three, five, and seven statistical boundary sets ($N = 800$) are also about a few percent for all diffraction orders under study. For the final efficiency modeling (Fig. 12.25), $N = 800$ and seven random boundary sets are used. The total error for all points and ranges derived from the energy balance (12.128)

was on the order of 10^{-3}. The time taken up by one rigorous computation (one scanning point) for $N = 800$ on the aforementioned workstation and operating system is ~ 45 min.

Convergence and accuracy of the scattering intensity results of a randomly-rough 41-boundary grating were investigated using similar accuracy parameters to the efficiency computation. A medium rate of convergence of the light intensity results was observed for a wavelength of 19.25 nm and the above computation model. More than 100 sets of 41 rough border profiles and medium number of discretization points are enough to compute exact values of scattering light intensities between orders. The difference between scattered light intensities data obtained with seven boundary sets using medium ($N = 1000$) and high ($N = 1200$) accuracy is about 10^{-5} for almost all diffraction angles. The differences between scattered light intensities obtained with different numbers of statistical boundary sets (35, 70, 98, 105) and $N = 1000$ for all diffraction (scattering) angles are shown in Ref. 12.1. For the final scattering intensity modeling (Fig. 12.26), $N = 1000$ and 105 random boundary sets were chosen. The total error for all points and ranges derived from the energy balance was on the order of 10^{-5}. The time taken up by one computation for $N = 1000$ on the aforementioned workstation and operating system is about two hours.

12.10 Appendix A: Derivation of the recursive algorithm for Separating solver

In any of the strips $\{u_j < y < d_{j-1}\}$ the functions ε and μ take constant values and we introduce its wave number κ_j by

$$\kappa_j^2 = \varepsilon\mu - \varepsilon_0\mu_0 \sin^2\phi\,.$$

As quasi-periodic solutions of the Helmholtz equation

$$(\Delta + \omega^2\kappa_j^2)\,u = 0$$

in the strips $\{u_j < y < d_{j-1}\}$ between Σ_j and Σ_{j-1}, $j = 1,\ldots,M-1$, the functions E_z, B_z are smooth and bounded. Hence, for $y \in (u_j, d_{j-1})$

$$(E_z, B_z) = \sum_{n\in\mathbb{Z}} \left((a_n^j, c_n^j)\,\mathrm{e}^{-i\beta_n^{(j)}y} + (b_n^j, d_n^j)\,\mathrm{e}^{i\beta_n^{(j)}y} \right) \mathrm{e}^{i\alpha_n x}\,.$$

Assign to each profile Σ_j a characteristic y-coordinate y_j, for example $y_j = Y_j(0)$ for a given parametrization $(X_j(t), Y_j(t))$ of the profile Σ_j. Recall that $y_0 > y_1 > \ldots > y_{M-1}$. Using the notation

$$\begin{aligned} &(A_n^j, C_n^j) = \mathrm{e}^{-i\beta_n^{(j)}y_j}(a_n^j, c_n^j)\,, \qquad (B_n^j, D_n^j) = \mathrm{e}^{i\beta_n^{(j)}y_j}(b_n^j, d_n^j)\,, \\ &(\mathscr{A}_n^j, \mathscr{C}_n^j) = \mathrm{e}^{-i\beta_n^{(j+1)}y_j}(a_n^{j+1}, c_n^{j+1})\,,\ (\mathscr{B}_n^j, \mathscr{D}_n^j) = \mathrm{e}^{i\beta_n^{(j+1)}y_j}(b_n^{j+1}, d_n^{j+1})\,, \end{aligned} \tag{12.141}$$

the field in $\{u_j < y < d_{j-1}\}$ above Σ_j is given by

$$(E_z, B_z) = \sum_{n\in\mathbb{Z}} \left((A_n^j, C_n^j)\,\mathrm{e}^{-i\beta_n^{(j)}(y-y_j)} + (B_n^j, D_n^j)\,\mathrm{e}^{i\beta_n^{(j)}(y-y_j)}] \right) \mathrm{e}^{i\alpha_n x}, \tag{12.142}$$

whereas in $\{u_{j+1} < y < d_j\}$ below Σ_j

$$(E_z, B_z) = \sum_{n\in\mathbb{Z}} \left((\mathscr{A}_n^j, \mathscr{C}_n^j)\,\mathrm{e}^{-i\beta_n^{(j+1)}(y-y_j)} + (\mathscr{B}_n^j, \mathscr{D}_n^j)\,\mathrm{e}^{i\beta_n^{(j+1)}(y-y_j)} \right) \mathrm{e}^{i\alpha_n x}. \tag{12.143}$$

The terms $(A_n^j, C_n^j)\,\mathrm{e}^{i\alpha_n x - i\beta_n^{(j)}(y-y_j)}$ and $(\mathscr{B}_n^j, \mathscr{D}_n^j)\,\mathrm{e}^{i\alpha_n x - i\beta_n^{(j+1)}(y-y_j)}$ correspond to incident waves on the profile Σ_j, whereas $(B_n^j, D_n^j)\,\mathrm{e}^{i\alpha_n x + i\beta_n^{(j)}(y-y_j)}$ and $(\mathscr{A}_n^j, \mathscr{C}_n^j)\,\mathrm{e}^{i\alpha_n x + i\beta_n^{(j+1)}(y-y_j)}$ represent the diffracted waves. Thus, the coefficients in equations (12.142) and (12.143) are linked by the reflection and transmission matrices of the grating having only the interface Σ_j.

For a compact notation, we introduce the infinite coefficient vectors

$$\mathbf{A}_j = (\ldots, A_{-1}^j, A_0^j, A_1^j, \ldots, C_{-1}^j, C_0^j, C_1^j, \ldots)^T,\ \mathscr{A}_j = (\ldots, \mathscr{A}_{-1}^j, \mathscr{A}_0^j, \mathscr{A}_1^j, \ldots, \mathscr{C}_{-1}^j, \mathscr{C}_0^j, \mathscr{C}_1^j, \ldots)^T.$$
$$\boldsymbol{B}_j = (\ldots, B_{-1}^j, B_0^j, B_1^j, \ldots, D_{-1}^j, D_0^j, D_1^j, \ldots)^T,\ \mathscr{B}_j = (\ldots, \mathscr{B}_{-1}^j, \mathscr{B}_0^j, \mathscr{B}_1^j, \ldots, \mathscr{D}_{-1}^j, \mathscr{D}_0^j, \mathscr{D}_1^j, \ldots)^T.$$

Then equations (12.141) can be written in the form

$$\mathscr{A}_{j-1} = \boldsymbol{\gamma}_j^{-1}\boldsymbol{A}_j\,,\ \mathscr{B}_{j-1} = \boldsymbol{\gamma}_j\boldsymbol{B}_j\,, \tag{12.144}$$

with the infinite diagonal matrix

$$\boldsymbol{\gamma}_j = \mathrm{diag}(\ldots, \mathrm{e}^{i\beta_{-1}^{(j)}h_j}, \mathrm{e}^{i\beta_0^{(j)}h_j}, \mathrm{e}^{i\beta_1^{(j)}h_j}, \ldots, \mathrm{e}^{i\beta_{-1}^{(j)}h_j}, \mathrm{e}^{i\beta_0^{(j)}h_j}, \mathrm{e}^{i\beta_1^{(j)}h_j}, \ldots),$$

with $h_j = y_{j-1} - y_j > 0$.

Denoting by $\mathbf{r}_j, \mathbf{t}_j$ the (infinite) reflection and transmission matrices of the grating with profile Σ_j for illumination from above and by $\mathbf{r}'_j, \mathbf{t}'_j$ the corresponding matrices for illumination of Σ_j from below. This means that the incoming field with coefficient vector $\mathbf{A}_j$ is diffracted by the simple grating with profile Σ_j into the reflected field with coefficient vector $\mathbf{r}_j\mathbf{A}_j$ and the transmitted field with coefficient vector $\mathbf{t}_j\mathbf{A}_j$. Analogously, illumination from below by a field with coefficient vector $\mathscr{B}_j$ results in a reflected field characterized by $\mathbf{r}'_j\mathscr{B}_j$ and a transmitted field with coefficient vector $\mathbf{t}'_j\mathscr{B}_j$. Hence, for any $j = 1, \ldots, M-2$ the coefficient vectors are linked by the relations

$$\boldsymbol{B}_j = \mathbf{r}_j\boldsymbol{A}_j + \mathbf{t}'_j\mathscr{B}_j\,,\quad \mathscr{A}_j = \mathbf{t}_j\boldsymbol{A}_j + \mathbf{r}'_j\mathscr{B}_j\,. \tag{12.145}$$

Writing (12.11) in the form

$$(E_z, B_z) = (A_0^0, C_0^0)\,\mathrm{e}^{i\alpha x - i\beta(y-y_0)} + \sum_{n\in\mathbb{Z}} (B_n^0, D_n^0)\,\mathrm{e}^{i\alpha_n x + i\beta_n^{(0)}(y-y_0)}\,.$$

we obtain (12.145) with $j = 0$, whereas for $y < -H$ we derive from

$$(E_z, B_z) = \sum_{n\in\mathbb{Z}} (\mathscr{A}_n^{M-1}, \mathscr{C}_n^{M-1})\,\mathrm{e}^{-i\beta_n^{(M)}(y-y_M)}\,\mathrm{e}^{i\alpha_n x}$$

the relation

$$\boldsymbol{B}_{M-1} = \mathbf{r}_{M-1}\boldsymbol{A}_{M-1}\,,\quad \mathscr{A}_{M-1} = \mathbf{t}_{M-1}\boldsymbol{A}_{M-1}\,. \tag{12.146}$$

Here we provide the formulas for solving the multi-profile problem to determine the vectors $\boldsymbol{B}_0$ and $\mathscr{A}_{M-1}$ from given input $\boldsymbol{A}_0$ and vanishing $\mathscr{B}_{M-1}$. The idea is to look for a recursion for the operators $\mathbf{R}_j, \mathbf{T}_j$ such that

$$\boldsymbol{B}_j = \mathbf{R}_j\boldsymbol{A}_j\,,\quad \mathscr{A}_{M-1} = \mathbf{T}_j\boldsymbol{A}_j\,,\quad j = M-1, \ldots, 0.$$

By (12.146) we know that $\mathbf{R}_{M-1} = \mathbf{r}_{M-1}$, $\mathbf{T}_{M-1} = \mathbf{t}_{M-1}$. Furthermore, we have from (12.144) and (12.145)

$$\mathbf{B}_{j-1} = \mathbf{r}_{j-1}\mathbf{A}_{j-1} + \mathbf{t}'_{j-1}\boldsymbol{\gamma}_j\mathbf{B}_j\,,\quad \boldsymbol{\gamma}_j^{-1}\mathbf{A}_j = \mathbf{t}_{j-1}\mathbf{A}_{j-1} + \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{B}_j\,,$$

which gives

$$\mathbf{B}_{j-1} = \mathbf{r}_{j-1}\mathbf{A}_{j-1} + \mathbf{t}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\mathbf{A}_j\,, \tag{12.147}$$

$$\boldsymbol{\gamma}_j^{-1}\mathbf{A}_j = \mathbf{t}_{j-1}\mathbf{A}_{j-1} + \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\mathbf{A}_j\,. \tag{12.148}$$

The last equation implies

$$\mathbf{A}_j = \left(\boldsymbol{\gamma}_j^{-1} - \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\mathbf{t}_{j-1}\mathbf{A}_{j-1}\,,$$

which transforms (12.147) into

$$\mathbf{B}_{j-1} = \left(\mathbf{r}_{j-1} + \mathbf{t}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\left(\boldsymbol{\gamma}_j^{-1} - \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\mathbf{t}_{j-1}\right)\mathbf{A}_{j-1}\,,$$

and hence

$$\mathbf{R}_{j-1} = \mathbf{r}_{j-1} + \mathbf{t}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\left(\boldsymbol{\gamma}_j^{-1} - \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\mathbf{t}_{j-1}\,. \tag{12.149}$$

Finally, from

$$\mathscr{A}_{N-1} = \mathbf{T}_j\left(\boldsymbol{\gamma}_j^{-1} - \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\mathbf{t}_{j-1}\mathbf{A}_{j-1}$$

we derive

$$\mathbf{T}_{j-1} = \mathbf{T}_j\left(\boldsymbol{\gamma}_j^{-1} - \mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\mathbf{t}_{j-1}\,, \tag{12.150}$$

This leads to the following marching procedure:

Set	$\mathbf{R}_{M-1} = \mathbf{r}_{M-1}\,,\ \mathbf{T}_{M-1} = \mathbf{t}_{M-1}\,;$
Compute for $j = M-1,\ldots,1$	$\mathbf{R}_{j-1} = \mathbf{r}_{j-1} + \mathbf{t}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\left(\mathbf{I} - \boldsymbol{\gamma}_j\mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\boldsymbol{\gamma}_j\mathbf{t}_{j-1}\,;$
	$\mathbf{T}_{j-1} = \mathbf{T}_j\left(\mathbf{I} - \boldsymbol{\gamma}_j\mathbf{r}'_{j-1}\boldsymbol{\gamma}_j\mathbf{R}_j\right)^{-1}\boldsymbol{\gamma}_j\mathbf{t}_{j-1}\,;$
Determine finally	$\boldsymbol{B}_0 = \mathbf{R}_0\boldsymbol{A}_0\,,\ \mathscr{A}_{M-1} = \mathbf{T}_0\boldsymbol{A}_0.$

12.11 Appendix B: Derivation of the recursive algorithm for Penetrating solver

The scheme is based on the ansatz

$$\begin{pmatrix} u_{j+1}|_{\Gamma_j} \\ v_{j+1}|_{\Gamma_j} \end{pmatrix} = \mathscr{A}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix}, \quad \begin{pmatrix} \partial_n u_{j+1}|_{\Gamma_j} \\ \partial_n v_{j+1}|_{\Gamma_j} \end{pmatrix} = \mathscr{B}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} \quad j = 0,\ldots,M-1\,, \tag{12.151}$$

with certain 2×2 linear operator matrices $\mathscr{A}_j$ and $\mathscr{B}_j$. Note first that the initial values (12.124) follow from (12.120) and the jump relation (12.26) for $\partial_n\mathscr{S}_{\Gamma_M-1,M}$.

Using (12.151) the transmission conditions (12.116) on Γ_j for $j = 1,\ldots,M-1$ can be written in the form

$$\begin{pmatrix} u_j|_{\Gamma_j} \\ v_j|_{\Gamma_j} \end{pmatrix} = \mathscr{A}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix}, \tag{12.152}$$

$$\begin{pmatrix} \partial_n u_j|_{\Gamma_j} \\ \partial_n v_j|_{\Gamma_j} \end{pmatrix} = \begin{pmatrix} a_j & 0 \\ 0 & b_j \end{pmatrix}\mathscr{B}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} + \begin{pmatrix} 0 & -c_j\,\partial_t \\ d_j\,\partial_t & 0 \end{pmatrix}\mathscr{A}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix}. \tag{12.153}$$

The representation (12.119) and the jump relation (12.26) of the double layer potential $\mathscr{D}_{\Gamma_j,j}$ imply that

$$u_j|_{\Gamma_j} = \frac{1}{2}\big(V_{jj}^{(j)}\partial_{\mathbf{n}}u_j - (K_{jj}^{(j)} - I)u_j\big) + V_{jj-1}^{(j)}\varphi_{j-1}\,,$$
$$v_j|_{\Gamma_j} = \frac{1}{2}\big(V_{jj}^{(j)}\partial_{\mathbf{n}}v_j - (K_{jj}^{(j)} - I)v_j\big) + V_{jj-1}^{(j)}\psi_{j-1}\,.$$

Hence (12.153) leads, in matrix notation, to the equation

$$\begin{pmatrix} a_jV_{jj}^{(j)} & 0 \\ 0 & b_jV_{jj}^{(j)} \end{pmatrix}\mathscr{B}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} - \begin{pmatrix} I+K_{jj}^{(j)} & c_jV_{jj}^{(j)}\partial_t \\ -d_jV_{jj}^{(j)}\partial_t & I+K_{jj}^{(j)}\end{pmatrix}\mathscr{A}_j\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} = -2\begin{pmatrix}V_{jj-1}^{(j)}\varphi_{j-1}\\ V_{jj-1}^{(j)}\psi_{j-1}\end{pmatrix}, \tag{12.154}$$

which is equivalent to (12.116). Using the singular integral $H_{jj}^{(j)} = -V_{jj}^{(j)}\partial_t$ (see (12.31)) we obtain the relation

$$\left(\begin{pmatrix} I+K_{jj}^{(j)} & -c_jH_{jj}^{(j)} \\ d_jH_{jj}^{(j)} & I+K_{jj}^{(j)}\end{pmatrix}\mathscr{A}_j - \begin{pmatrix} a_jV_{jj}^{(j)} & 0 \\ 0 & b_jV_{jj}^{(j)}\end{pmatrix}\mathscr{B}_j\right)\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} = 2\begin{pmatrix} V_{jj-1}^{(j)} & 0 \\ 0 & V_{jj-1}^{(j)}\end{pmatrix}\begin{pmatrix}\varphi_{j-1}\\ \psi_{j-1}\end{pmatrix},$$

which is satisfied by

$$\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix} = \mathscr{Q}_{j-1}\begin{pmatrix}\varphi_{j-1}\\ \psi_{j-1}\end{pmatrix},$$

provided that $\mathscr{Q}_{j-1}$ is a solution of the operator equation (12.123).

The equations (12.125) and (12.126) for $\mathscr{A}_{j-1}$ and $\mathscr{B}_{j-1}$ are derived from relations on the upper boundary Γ_{j-1} of G_j. The representation (12.119) and condition (12.153) give

$$\begin{aligned}\begin{pmatrix}u_j|_{\Gamma_{j-1}}\\ v_j|_{\Gamma_{j-1}}\end{pmatrix} &= \frac{1}{2}\left(\begin{pmatrix} V_{j-1j}^{(j)} & 0\\ 0 & V_{j-1j}^{(j)}\end{pmatrix}\begin{pmatrix}\partial_n u_j|_{\Gamma_j}\\ \partial_n v_j|_{\Gamma_j}\end{pmatrix} - \begin{pmatrix} K_{j-1j}^{(j)} & 0\\ 0 & K_{j-1j}^{(j)}\end{pmatrix}\begin{pmatrix}u_j|_{\Gamma_j}\\ v_j|_{\Gamma_j}\end{pmatrix}\right) + \begin{pmatrix}V_{j-1j-1}^{(j)}\varphi_{j-1}\\ V_{j-1j-1}^{(j)}\psi_{j-1}\end{pmatrix}\\ &= \frac{1}{2}\left(\begin{pmatrix} a_jV_{j-1j}^{(j)} & 0\\ 0 & b_jV_{j-1j}^{(j)}\end{pmatrix}\mathscr{B}_j - \begin{pmatrix} K_{j-1j}^{(j)} & c_jV_{j-1j}^{(j)}\partial_t\\ -d_jV_{j-1j}^{(j)}\partial_t & K_{j-1j}^{(j)}\end{pmatrix}\mathscr{A}_j\right)\begin{pmatrix}\varphi_j\\ \psi_j\end{pmatrix}\\ &\quad + \begin{pmatrix} V_{j-1j-1}^{(j)} & 0\\ 0 & V_{j-1j-1}^{(j)}\end{pmatrix}\begin{pmatrix}\varphi_{j-1}\\ \psi_{j-1}\end{pmatrix},\end{aligned}$$

which by (12.151), (12.121) and using $H_{j-1j}^{(j)} = -V_{j-1j}^{(j)}\partial_t$ leads to (12.125).

Now (12.126) follows from (12.23) and (12.119), since

$$V_{j-1j-1}^{(j)}\varphi_{j-1} = -\frac{1}{2}\big(V_{j-1j-1}^{(j)}\partial_n u_j - (I+K_{j-1j-1}^{(j)})u_j\big)\,,$$
$$V_{j-1j-1}^{(j)}\psi_{j-1} = -\frac{1}{2}\big(V_{j-1j-1}^{(j)}\partial_n v_j - (I+K_{j-1j-1}^{(j)})v_j\big)\,,$$

imply that on Γ_{j-1}

$$\begin{pmatrix}\partial_n u_j\\ \partial_n v_j\end{pmatrix} = \begin{pmatrix}(V_{j-1j-1}^{(j)})^{-1}(I+K_{j-1j-1}^{(j)}) & 0\\ 0 & (V_{j-1j-1}^{(j)})^{-1}(I+K_{j-1j-1}^{(j)})\end{pmatrix}\mathscr{A}_{j-1}\begin{pmatrix}\varphi_{j-1}\\ \psi_{j-1}\end{pmatrix} - 2\begin{pmatrix}\varphi_{j-1}\\ \psi_{j-1}\end{pmatrix}.$$

Equation (12.127) follows from the relations

$$V_{00}^{(0)}\,\partial_n E_z^i - (I+K_{00}^{(0)})E_z^i = -2E_z^i, \quad V_{00}^{(0)}\,\partial_n B_z^i - (I+K_{00}^{(0)})B_z^i = -2B_z^i$$

on the upper profile Γ_0, which hold because E_z^i, B_z^i satisfy the Helmholtz equation $(\Delta+\omega^2\kappa_0^2)u = 0$ and the outgoing wave condition in $G_0^- = \mathbb{R}^2 \setminus \overline{G_0}$. Hence, the transmission conditions (12.115) are fulfilled if and only if

$$\begin{pmatrix} I+K_{00}^{(0)} & -c_0 H_{00}^{(0)} \\ d_0 H_{00}^{(0)} & I+K_{00}^{(0)} \end{pmatrix} \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} - \begin{pmatrix} a_0 V_{00}^{(0)} & 0 \\ 0 & b_0 V_{00}^{(0)} \end{pmatrix} \begin{pmatrix} \partial_n u_1 \\ \partial_n v_1 \end{pmatrix} = -2 \begin{pmatrix} u^i \\ v^i \end{pmatrix},$$

i.e., if φ_0, ψ_0 satisfy (12.127).

Remark 12.11.1 *If the material in the bottom layer G_M is a perfect conductor, then the z-components of E and B have to satisfy the boundary condition*

$$E_z = u_M = 0\,, \quad \partial_n B_z = \partial_n v_M = 0 \qquad \text{on } \Gamma_{M-1}\,. \tag{12.155}$$

In this case, it is easy to see that the relations (12.125) *and* (12.126) *for $j = M-1$ with the coefficients $a_{M-1} = 1$, $b_{M-1} = c_{M-1} = d_{M-1} = 0$, and the initial values*

$$\mathscr{A}_{M-1} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \text{ and } \quad \mathscr{B}_{M-1} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

lead to $\mathscr{A}_{M-2}$ and $\mathscr{B}_{M-2}$ satisfying

$$\begin{pmatrix} u_{M-1}|_{\Gamma_{M-2}} \\ v_{M-1}|_{\Gamma_{M-2}} \end{pmatrix} = \mathscr{A}_{M-2} \begin{pmatrix} \varphi_{M-2} \\ \psi_{M-2} \end{pmatrix}, \; \begin{pmatrix} \partial_n u_{M-1}|_{\Gamma_{M-2}} \\ \partial_n v_{M-1}|_{\Gamma_{M-2}} \end{pmatrix} = \mathscr{B}_{M-2} \begin{pmatrix} \varphi_{M-2} \\ \psi_{M-2} \end{pmatrix}.$$

Hence, the densities $\{\varphi_j, \psi_j\}$, $j = 0, \ldots, M-2$, are derived by the same scheme (12.121 - 12.127).

12.12 Appendix C: Derivation of the absorption energy for multilayer gratings

As in Section 12.3.3 the application of Helmholtz equations and Green's formula in $\Omega_H \cap G_0$ implies the relation

$$\frac{\varepsilon_0}{\varepsilon_v}|p_z|^2 + \frac{\mu_0}{\mu_v}|q_z|^2 = \sum_{\beta_n^0 \geq 0} \frac{\beta_n^0}{\beta} \Big(\frac{\varepsilon_0}{\varepsilon_v}|E_n^0|^2 + \frac{\mu_0}{\mu_v}|B_n^0|^2 \Big) + \frac{\varepsilon_0}{\varepsilon_v \beta} \operatorname{Im} \int_{\Gamma_0} \partial_n E_z \overline{E_z} + \frac{\mu_0}{\mu_v \beta} \operatorname{Im} \int_{\Gamma_0} \partial_n B_z \overline{B_z}\,.$$

where (E_z, B_z) is the solution of the conical diffraction problem, and $\partial_n E_z = \partial_n^+ E_z$, $\partial_n B_z = \partial_n^+ B_z$ are the normal derivatives on Γ_0 of the z-components of the total fields in G_0, i.e. the sum of the reflected and the incident fields. Setting the energy of the incident wave

$$\frac{\varepsilon_0}{\varepsilon_v}|p_z|^2 + \frac{\mu_0}{\mu_v}|q_z|^2 = 1\,,$$

the sum of reflection order efficiencies R (cf. (12.67)) fulfils

$$R + \frac{\varepsilon_0}{\varepsilon_v \beta} \operatorname{Im} \int_{\Gamma_0} \partial_n E_z \overline{E_z} + \frac{\mu_0}{\mu_v \beta} \operatorname{Im} \int_{\Gamma_0} \partial_n B_z \overline{B_z} = 1\,.$$

Hence, for non-transparent grating we derive the energy conservation

$$R+A=1$$

with the absorption A

$$A=\frac{\varepsilon_0}{\varepsilon_v\beta}\,\mathrm{Im}\int_{\Gamma_0}\partial_n E_z\overline{E_z}+\frac{\mu_0}{\mu_v\beta}\,\mathrm{Im}\int_{\Gamma_0}\partial_n B_z\overline{B_z}. \tag{12.156}$$

If, otherwise, the material parameters ε_M and μ_M are real, then some part of the incident field will be transmitted. Then similar considerations in the domain $\Omega_H\cap G_M$ lead to the relations

$$T-\frac{\varepsilon_M\kappa_0^2}{\varepsilon_v\beta\,\kappa_M^2}\,\mathrm{Im}\int_{\Gamma_{M-1}}\partial_n E_z\overline{E_z}-\frac{\mu_M\kappa_0^2}{\mu_v\beta\,\kappa_M^2}\,\mathrm{Im}\int_{\Gamma_{M-1}}\partial_n B_z\overline{B_z}=0\,,$$

where T is the sum of transmission order efficiencies of the multilayer grating (cf. (12.68)), and $\partial_n E_z=\partial_n^- E_z$, $\partial_n B_z=\partial_n^- B_z$ are the normal derivatives on Γ_{M-1} of the z-components of the transmitted fields in G_M. In this case we derive the energy conservation

$$R+T+A=1\,,$$

where the absorption A is given by the formula

$$A=\frac{1}{\beta}\,\mathrm{Im}\int_{\Gamma_0}\Big(\frac{\varepsilon_0}{\varepsilon_v}\partial_n E_z\overline{E_z}+\frac{\mu_0}{\mu_v\beta}\partial_n B_z\overline{B_z}\Big)-\frac{\kappa_0^2}{\beta\,\kappa_M^2}\,\mathrm{Im}\int_{\Gamma_{M-1}}\Big(\frac{\varepsilon_M}{\varepsilon_v}\partial_n E_z\overline{E_z}+\frac{\mu_M}{\mu_v}\partial_n B_z\overline{B_z}\Big). \tag{12.157}$$

Using the jump conditions the obtained formulas for A can be easily transformed. For example, from (12.115) we know that on Γ_0

$$\begin{aligned}
&\frac{\varepsilon_1\,\partial_n^- E_z}{\varepsilon_v\kappa_1^2}-\frac{\varepsilon_0\partial_n^+ E_z}{\varepsilon_v\kappa_0^2}=\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_0^2}-\frac{1}{\kappa_1^2}\Big)\partial_t B_z\,,\\
&\frac{\mu_1\partial_n^- B_z}{\mu_v\kappa_1^2}-\frac{\mu_0\partial_n^+ B_z}{\mu_v\kappa_0^2}=-\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_0^2}-\frac{1}{\kappa_1^2}\Big)\partial_t E_z\,,
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathrm{Im}\int_{\Gamma_0}\Big(\frac{\varepsilon_0}{\varepsilon_v}\partial_n^+ E_z\overline{E_z}+\frac{\mu_0}{\mu_v\beta}\partial_n^+ B_z\overline{B_z}\Big)\\
&=\kappa_0^2\Big(\mathrm{Im}\int_{\Gamma_0}\frac{1}{\kappa_1^2}\Big(\frac{\varepsilon_1}{\varepsilon_v}\partial_n^- E_z\overline{E_z}+\frac{\mu_1}{\mu_v}\partial_n^- B_z\overline{B_z}\Big)+\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\,\mathrm{Im}\,\frac{2\sin\phi}{\kappa_1^2}\,\mathrm{Re}\int_{\Gamma_0}E_z\overline{\partial_t B_z}\Big).
\end{aligned}$$

Further, on Γ_{M-1}

$$\begin{aligned}
&\frac{\varepsilon_M\,\partial_n^- E_z}{\varepsilon_v\kappa_M^2}-\frac{\varepsilon_{M-1}\partial_n^+ E_z}{\varepsilon_v\kappa_{M-1}^2}=\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_{M-1}^2}-\frac{1}{\kappa_M^2}\Big)\partial_t B_z\,,\\
&\frac{\mu_M\partial_n^- B_z}{\mu_v\kappa_M^2}-\frac{\mu_{M-1}\partial_n^+ B_z}{\mu_v\kappa_{M-1}^2}=-\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\sin\phi\,\Big(\frac{1}{\kappa_{M-1}^2}-\frac{1}{\kappa_M^2}\Big)\partial_t E_z\,,
\end{aligned}$$

such that

$$\begin{aligned}
&\frac{\kappa_0^2}{\kappa_M^2}\,\mathrm{Im}\int_{\Gamma_{M-1}}\Big(\frac{\varepsilon_M}{\varepsilon_v}\partial_n^- E_z\overline{E_z}+\frac{\mu_M}{\mu_v}\partial_n^- B_z\overline{B_z}\Big)\\
&=\kappa_0^2\Big(\mathrm{Im}\int_{\Gamma_{M-1}}\frac{1}{\kappa_{M-1}^2}\Big(\frac{\varepsilon_{M-1}}{\varepsilon_v}\partial_n^+ E_z\overline{E_z}+\frac{\mu_{M-1}}{\mu_v}\partial_n^+ B_z\overline{B_z}\Big)+\sqrt{\frac{\varepsilon_0\mu_0}{\varepsilon_v\mu_v}}\,\mathrm{Im}\,\frac{2\sin\phi}{\kappa_{M-1}^2}\,\mathrm{Re}\int_{\Gamma_{M-1}}E_z\overline{\partial_t B_z}\Big).
\end{aligned}$$

12.13 Appendix D: Derivation of the general connection rule between 2D and 1D gratings

We seek for a perturbative development of the reflection operator **R** in powers of the heights $h_x^{(i)}$ and $h_z^{(j)}$ of a bi-periodic surface either conductive or dielectric that is the sum of the two Fourier series:

$$\begin{aligned} h(x,z) = h_x + h_z = \sum_i h_x^{(i)} \sin(2\pi x i/d_x + \tau_x^{(i)}) \\ + \sum_j h_z^{(j)} \sin(2\pi z j/d_z + \tau_z^{(j)}) \end{aligned} \tag{12.158}$$

Such a representation of $h(x,z)$ is typical for real 2D periodic or random surfaces obtained, e.g., as a linear response of a photoresist to light with two separate exposures in perpendicular planes or by polishing using a linear tool. Note that the 2D Fourier transformation of $h(x,z)$ is also the sum of two 1D Fourier transforms of h_x and h_z. We suppose also that the bigrating works under arbitrary incidence and polarization states of a plane monochromatic wave and the respective single-periodic gratings work in conical diffraction. Suppose for simplicity h_x and h_z are even functions, which is true for many ergodic stationary processes. So, replacing h_x or h_z by $-h_x$ or $-h_z$ does not change the diffraction pattern in the far-field zone. We will study the perturbative expansion of the reflected efficiency η as a function of the surface heights $(h_x^{(i)})^2$ and $(h_z^{(j)})^2$. Using the perturbative expansion of **R**, the terms of η which contain an expression such as $(h_x^{(i)})^{2k}$, $(h_z^{(j)})^{2l}$ will be denoted R_{kl}:

$$\eta = R_{00} + R_{01} + R_{10} + R_{11} + R_{02} + R_{20} + \ldots$$

Using the quasi-periodicity property of **R** and Taylor expansion of scattered field amplitudes in powers of the surface profile heights (e.g, see Eq. 53 of Ref. 12.50), η_{mn}, η_m, and η_n can be expressed in the following forms:

$$\begin{aligned} \eta_{mn} - o(h^6) = \delta_{mn} a_{00} + \sum_i a_{10}^{(i)} (h_x^{(i)})^2 + a_{20}^{(i)} (h_x^{(i)})^4 \\ + \sum_j a_{01}^{(j)} (h_z^{(j)})^2 + a_{02}^{(j)} (h_z^{(j)})^4 + \sum_{i,j} a_{11}^{(i,j)} (h_x^{(i)} h_z^{(j)})^2, \end{aligned} \tag{12.159}$$

$$\eta_m - o(h_x^6) = \delta_{m0} a_{00} + \sum_i a_{10}^{(i)} (h_x^{(i)})^2 + a_{20}^{(i)} (h_x^{(i)})^4, \tag{12.160}$$

$$\eta_n - o(h_z^6) = \delta_{0n} a_{00} + \sum_j a_{01}^{(j)} (h_z^{(j)})^2 + a_{02}^{(j)} (h_z^{(j)})^4, \tag{12.161}$$

where $\delta_{m,n}$ is the Kronecker delta.

From (12.159)–(12.161) we choose from the physical point of view one of the two possible expressions for η_{mn} through η_m and η_n:

$$\eta_{mn} - o(h^6) = \frac{\eta_m \eta_n}{a_{00}} + \sum_{i,j} (a_{11}^{(i,j)} - \frac{a_{m0}^{(i)} a_{0n}^{(j)}}{a_{00}}) (h_x^{(i)} h_z^{(j)})^2 \tag{12.162}$$

Finally, using (12.162) one can formulate the equivalence rule:

$$\eta_{mn} = \frac{\eta_m \eta_n}{r_F} + o(h^4), m \vee n = 0, h_{x,z}/d_{x,z} < 1, \tag{12.163}$$

where η_m and η_n are 1D grating efficiencies obtained in conical diffraction, r_F—the Fresnel factor of a 2D surface. It is worth noting that η_m and η_n in this equivalence rule should be computed with preservation of incidence and polarization angles of both gratings in the absolute coordinate system.

References:

[1] [http://www.pcgrate.com/] (2014).

[2] [http://www.wias-berlin.de/software/DIPOG/?lang=1/] (2014).

[3] E. G. Loewen and E. Popov, *Diffraction Gratings and Applications* (Marcel Dekker, New York, 1997).

[4] L. I. Goray, J. F. Seely, and S. Yu. Sadov, "Spectral separation of the efficiencies of the inside and outside orders of soft-x-ray-extreme-ultraviolet gratings at near normal incidence," J. Appl. Phys. **100**, 094901-1–13 (2006).

[5] L. I. Goray, "Application of the boundary integral equation method to very small wavelength-to-period diffraction problems," Waves Random Media **20**, 569-586 (2010).

[6] L. I. Goray, "Application of the rigorous method to x-ray and neutron beam scattering on rough surfaces," J. Appl. Phys. **108**, 033516-1–10 (2010).

[7] L. I. Goray and G. Schmidt, "Solving conical diffraction grating problems with integral equations," J. Opt. Soc. Am. A **27**, 585-597 (2010).

[8] Y. Wu and Y. Y. Lu, "Boundary integral equation Neumann-to-Dirichlet map method for gratings in conical diffraction," J. Opt. Soc. Am. A **28**, 1191-1196 (2011).

[9] G. Schmidt, "On the Diffraction by Biperiodic Anisotropic Structures," Appl. Anal. **82**, 75-92 (2010)

[10] J. Seely, B. Kjornrattanawanich, L. Goray, Y. Feng, and J. Bremer, "Characterization of zone plate properties using monochromatic synchrotron radiation in the 2 to 20 nm wavelength range," Appl. Opt. **50**, 3015-3020 (2011).

[11] G. Schmidt and B. H. Kleemann, "Integral equation methods from grating theory to photonics: an overview and new approaches for conical diffraction," J. Mod. Opt. **58**, 407-423 (2011).

[12] D. Maystre, "Photonic crystal diffraction gratings," Optics Express **8**, 209-216 (2001).

[13] L. I. Goray and G. Schmidt, "Analysis of two-dimensional photonic band gaps of any rod shape and conductivity using a conical-integral-equation method," Phys. Rev. E. **85**, 036701-1–12 (2012).

[14] B. Gallinet, A. M. Kern, and O. J. F. Martin, "Accurate and versatile modeling of electromagnetic scattering on periodic nanostructures with a surface integral approach," J. Opt. Soc. Am. A **27**, 2261-2271 (2010).

[15] V. Yu. Gotlib, "On solutions of the Helmholtz equation that are concentrated near a plane periodic boundary," J. Math. Sci. **102**, 41884194 (2000).

[16] G. Schmidt, "Boundary integral methods for periodic scattering problems," in *Around the Research of Vladimir Maz'ya II. Partial Differential Equations* (A. Laptev, ed., Springer, 2010), 337-363.

[17] R. Kress, "Boundary integral equations in time harmonics scattering," Math. Comput. Modelling **15**, 229-243 (1991).

[18] R. Kress, "A Nyström method for boundary integral equations in domains with corners," Num. Math. **58**, 145-161 (1990).

[19] O. P. Bruno, J. S. Ovall, and C. Turc,, A high-order integral algorithm for highly singular PDE solutions in Lipschitz domains," Computing **84**, 149-181 (2009).

[20] J. Bremer, "On the Nystrom discretization of integral equations on planar curves with corners," Applied and Computational Harmonic Analysis **32**, 45-64 (2012).

[21] A. Pomp, "The integral method for coated gratings: computational cost," **38**, 109-120 (1991).

[22] A. Rathsfeld, G. Schmidt, and B. H. Kleemann, "On a Fast Integral Equation Method for Diffraction GratingsCommun. Comput. Phys. **1**, 984-1009 (2006).

[23] C. M. Linton, "The Greens function for the two-dimensional Helmholtz equation in periodic domains," J. Eng. Math., 33 (1998), 377402.

[24] S. Yu. Sadov, "Computation of quasiperiodic fundamental solution of Helmholtz equation," in *Advances in Difference Equations* (I. Gyori, G. Ladas, and S. Elaydi, eds., Gordon and Breach, 1997), 551558.

[25] L. I. Goray and S. Yu. Sadov, Numerical modeling of coated gratings in sensetive cases," OSA Trends in Optics and Photonics Series **75**, 365-378 (2002).

[26] I. A. Abramowitz and M. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (New York, Dover Publications, 1972).

[27] B. Kleemann, A. Mitreiter, and F. Wyrowski, "Integral equati on m ethod with parametrization of grating profile Theory and experiments," J. Mod. Opt. **43**, 1323-1349 (1996).

[28] K. E. Atkinson, *The numerical solution of integral equations of the second kind* (Cambridge Univ. Press, Cambridge, 1997).

[29] D. Knuth, *The Art of Computer Programming, v. 2* (Addison-Wesley, 1968/1976/1998).

[30] L. Li, "Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings," J. Opt. Soc. Am. A **13**, 1024-1035 (1996).

[31] D. Maystre, "Electromagnetic study of photonic band gaps," Pur. Appl. Opt. **3**, 975-993 (1994).

[32] D. Maystre, "A new general integral theory for dielectric coated gratings", J. Opt. Soc. Am. **68**, 490-495 (1978).

[33] L.I. Goray, *Proc. of the Int. Conf. Days on Diffraction 2012*, IEEE, 98-103 (2012).

[34] A. Aho, J. Hopcroft, J. Ullman, *The Design and Analysis of Computer Algorithms* (Addisson-Wesley, 1976).

[35] L. I. Goray, "Numerical analysis of the efficiency of multilayer-coated gratings using integral method," Nucl. Instrum. Methods Phys. Res. A **536**, 211-221 (2005).

[36] E. D. Palik, ed., *Handbook of Optical Constant of Solids* (Academic, Orlando, 1985).

[37] R. Petit, ed., *Electromagnetic theory of gratings* (Springer, Berlin, 1980).

[38] L. Tsang, J. A. Kong, K.-H. Ding, C. O. Ao, *Scattering of Electromagnetics Waves: Numerical Simulations* (Wiley, New York, 2001).

[39] K. F. Warnick and W. C. Chew, Numerical simulation methods for rough surface scattering," Waves Random Media **11**, R1-R30 (2001).

[40] U. Pietsch, V. Holy, and T. Baumbach, *High-Resolution X-Ray Scattering: From Thin Films to Lateral Nanostructures* (Springer-Verlag, Heidelberg, 2004).

[41] T. M. Elfouhaily and C.-A. Guerin, "A critical survey of approximate scattering wave theories from random rough surfaces," Waves Random Media **14**, R1-R40 (2004).

[42] A. A. Maradudin, ed., *Light Scattering and Nanoscale Surface Roughness* (Springer, New York, 2007).

[43] D. Maystre and J. C. Dainty, eds., *Modern Analysis of Scattering Phenomena* (Hilger, New York, 1991).

[44] M. Nieto-Vesperinas and J. C. Dainty, eds., *Scattering in Volumes and Surfaces* (Elsevier, North-Holland, 1990).

[45] L. I. Goray, N. I. Chkhalo, and G. E. Tsyrlin, "Determining Angles of Incidence and Heights of Quantum Dot Faces by Analyzing X-ray Diffuse and Specular Scattering," Technical Physics **54**, 561-568 (2009).

[46] D. L. Voronov, E. H. Anderson, R. Cambie, S. Cabrini, S. D. Dhuey, L. I. Goray, E. M. Gullikson, F. Salmassi, T. Warwick, V. V. Yashchuk, and H. A. Padmore, "A 10,000 groove/mm multilayer coated grating for EUV spectroscopy," Opt. Express **19**, 6320-6325 (2011).

[47] L. Goray and M. Lubov, "Nonlinear continuum growth model of multiscale reliefs as applied to rigorous analysis of multilayer short-wave scattering intensity. I. Gratings," J. Appl. Cryst. **46**, 926-932 (2013).

[48] J. A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces* (IOP Publishing, Bristol, 1991).

[49] M. Saillard and A. Sentenac, "Rigorous solutions for electromagnetic scattering from rough surfaces," Waves Random Media **11**, R103-R137 (2001).

[50] A. Soubret , G. Berginc, andC. Bourrely, "Application of reduced Rayleigh equations to electromagnetic wave scattering by two-dimensional randomly rough surfaces," Phys. Rev. B **63**, 245411-1–20 (2001).

[51] J. B. Harris, T. W. Preist, J. R. Sambles, R. N. Thorpe, andR. A. Watts, "Optical response of bigratings," J. Opt. Soc. Am. A **13**, 2041-2049 (1996).

[52] L. I. Goray, I. G. Kuznetsov, S. Yu. Sadov, and D. A. Content, "Multilayer resonant subwavelength gratings: effects of waveguide modes and real groove profiles," J. Opt. Soc. Am. A **23**, 155-165 (2006).

[53] [http://http://ixo.gsfc.nasa.gov/technology/xgs.html] (2013).

[54] D. L. Voronov, E. H. Anderson, R. Cambie, P. Gawlitza, L. I. Goray, E. M. Gullikson, F. Salmassi, T. Warwick, V. V. Yashchuk, and H. A. Padmore, "Development of near atomically perfect diffraction gratings for EUV and soft x-rays with very high efficiency and resolving power," J. of Phys. Series C **25**, 152006-1–4 (2013).

[55] [http://henke.lbl.gov/optical/constants/] (2013).

[56] [http://hinode.nao.ac.jp/eise/] (2013).

[57] C. Tarrio, R. Watts, T. Lucatorto, J. Slaughter, and C. Falco, “Optical Constants of In Situ-Deposited Films of Important Extreme-Ultraviolet Multilayer Mirror Materials,” Appl. Opt. **37**, 4100-4104 (1998).

Chapter 13:

Fourier Modal Method

Lifeng Li

Table of Contents

Chapter 13

Fourier Modal Method

Lifeng Li

Department of Precision Instrument, Tsinghua University
Beijing 100084, China
lifengli@tsinghua.edu.cn

13.1 Introduction

The Fourier modal method is the most popular method for modeling diffraction gratings. The method is characterized by expanding the electromagnetic fields into Floquet-Fourier series and medium permittivity (and possibly also medium permeability) into Fourier series and solving the resulting Maxwell equations as a matrix eigenvalue problem. It is favored by researchers, engineers and graduate students who want to solve their research or engineering problems quickly. The popularity of the method is mostly due to its simplicity and partially due to its versatility. Compared with the other methods described in this book, to implement the Fourier modal method in a high-level computer language requires very few steps in programming and a minimal level of understanding of the mathematical theory behind it. The method is also fairly versatile because it can model both surface-relief gratings and volume gratings, i.e., gratings consisting of distinct periodic material boundaries and gratings having continuous periodic spatial variation of index of refraction. For a surface-relief grating of arbitrary shape, the method uses staircase approximation to approximate the profile function. When the periodic region of the grating contains only dielectric media or when the region contains metallic media but the electric field vector is everywhere parallel to the metal surfaces, the staircase approximation can produce reliable numerical results for the far field; however, when the latter condition is not met the approximation may fail to produce correct results.

There have been a couple of confusion points about the Fourier modal method. The first concerns its relationship with the so-called rigorous coupled-wave analysis. The latter name was first adopted by Moharam and Gaylord [13.1] in 1981 and it stemmed from the work of Kogelnik [13.2] in 1969 on an approximate theory for thick hologram gratings. Some people mistakenly consider that the two names refer to two related but different methods, whereas the two refer to exactly the same method since they solve the same modal equation in Fourier space. The name Fourier modal method is a better one because it reflects the essence of the method. The second confusion point is about the history of the method. Many people have been misled to believe that the method was proposed in the early 1980s [13.3, 13.4]. This is very unfortunate because the Fourier modal method has a history that is equally long as the classical differential method and integral method. For a brief review of its history the reader may consult with the introduction in [13.5].

Despite its long history and popularity, to date the Fourier modal method has not been given a complete and concise coverage in the literature. This is especially true for one-dimensional gratings in conical mounting. This chapter is an attempt to fill the void. The em-

phasis of this writing is on completeness and clarity of the formulation, and the aim is to help a reader correctly and easily implement the method in a computer code, if there is such a need.

To reach this goal within a reasonable number of pages I will cover only gratings whose surface profiles completely coincide with some Cartesian coordinate surfaces, although to achieve maximum generality without too much complication the coordinate system is assumed to be oblique. The technique of adaptive spatial resolution [13.6] and the method of matched coordinates [13.7] will not be described. To cover them it is necessary to blur the boundary between the Fourier modal method and the coordinate transformation method, but it is beyond the scope of this chapter. The Fourier factorization rules [13.5, 13.8] are assumed to be known to the reader and they will not be treated in great details. The validity and convergence rate of the Fourier modal method in the rigorous mathematical sense has rarely been addressed [13.5, 13.9] and they will not be discussed here. However, from a practical point of view the reader can be assured that the validity and accuracy of the method has been proven beyond doubt by numerous numerical tests and experimental verifications.

To make this chapter easier to read for a novice I have taken an approach of gradually increasing the difficulty of contents and brevity of description. Each section is built upon the previous one and extends the previous one in certain directions.

13.2 One-dimensional, isotropic gratings in non-conical mounting

13.2.1 General methodology

In the most general term the electromagnetic grating problem can be stated as this: given a periodic grating with known opto-geometric parameters and a monochromatic incident plane wave to find the distribution of the diffracted electromagnetic waves in the far field and near field. From an electromagnetic point of view this amounts to solving a boundary value problem. The Fourier modal method is just one way to solve this boundary value problem. The method consists of three steps. The three-dimensional physical space containing the grating is divided into three regions: the top semi-infinite transparent region where the incident plane wave comes from infinity, the middle region, also called the grating region or periodic region, where periodic variation of medium boundary or refractive indices takes place, and the bottom semi-infinite region that may be opaque. In some cases the middle region may be further divided into a few sub-regions. In the first step general expressions of the total electromagnetic fields in the individual regions (and sub-regions, if any) are found. These expressions still contain unknown coefficients. In the second step the electromagnetic boundary conditions are applied to the total fields at the interfaces between the regions and sub-regions, and the unknown coefficients are thereby determined by solving the resulting linear system of equations. In the final step the quantities of interests, e.g., diffraction efficiency, diffraction phase, state of polarization, and field distribution, etc., are extracted from the determined coefficients.

The three steps described above are actually shared by many numerical grating methods. It is the contents of the first step that distinguish one method from another. In a modal method the construction of the total electromagnetic fields is accomplished by means of modes. Mode is a powerful concept in physics. It means a distinct, self-sustainable pattern of motion that satisfies the governing law of physics including the internal boundary conditions. Just like a vibrating string has its vibrating mechanical or acoustic modes and an optical fiber has its transmission modes, a grating structure also has its electromagnetic modes. The modes of each region satisfy Maxwell equations and the associated internal boundary conditions, including the pseudo-periodicity conditions. In a modal method the solution of the total fields that satisfy the external boundary conditions between different regions and the radiation conditions at infinity is formed by superposition of all the modes, thanks to the linearity of the electromagnetic problem. In the two semi-infinite regions the total fields are given by the

Rayleigh expansions (see Chap. 2), and each term of the expansions can be rightfully viewed as a mode of the region.

The modes of the periodic region can be found in several ways. When the medium parameters and the electromagnetic fields are expanded into Fourier and Floquet-Fourier series, respectively, the resulting method of solution is called the Fourier modal method. Although in principle gratings of any groove shape possess modes, only the rectangular grating allows its modes to be easily found and lends itself as the basic object of study to the Fourier modal method. Another good reason for the Fourier modal method to be based on treatment of rectangular gratings is that a rectangle (or in general a parallelogram) is naturally a brick for building up an arbitrary grating profile (see Sect. 13.5).

In this section the Fourier modal method for the simplest grating problem, that of a rectangular grating in nonconical mounting, is formulated. Besides giving results for this simple but practically important grating case, the section also lays down the general framework for the remaining sections on more general grating shapes and plane wave incident angles.

13.2.2 Formulation in the rectangular Cartesian coordinate system

13.2.2.1 Description of the problem of rectangular gratings

Figure 13.1 depicts a rectangular grating illuminated by a monochromatic plane wave. A rectangular Cartesian coordinate system is attached to the grating with its x and y axes perpendicular and parallel to grating grooves, respectively. The xz plane is called the principal plane of the grating and the xy plane is called grating plane. The origin of the coordinate system is placed at the bottom center of one of the grooves. The grating is composed of at least two media, medium a and medium b with scalar permittivities ε_a and ε_b. The upper semi-infinite region (the cover) and the lower semi-infinite region (the substrate) are labeled with 2 and 0, and their scalar permittivities are $\varepsilon^{(2)}$ and $\varepsilon^{(0)}$. The magnetic permeabilities of all media are denoted by μ with appropriate subscripts or superscripts. Although in optical problems they are equal to that of vacuum the notation is formally retained for electromagnetic symmetry considerations. In most applications the cover is air and $\varepsilon_a = \varepsilon^{(2)}$, and very often the substrate and medium b are of the same material ($\varepsilon_b = \varepsilon^{(0)}$), so the spaces occupied by media a and b are called grooves and ridges of the grating, respectively. The grating period, groove depth, groove width, and ridge width are denoted by d, h, d_1, and w, respectively. Instead of w or d_1, the ratio w/d that is called the duty cycle (or filling factor) is often a more convenient parameter to use. For later convenience, we define two real ordinates, $z_0 = 0$ and $z_1 = h$, which are the lower and upper boundaries of the periodic layer, and two fictitious ordinates $z_{-1} = z_0$ and $z_2 = z_1$, which have no physical meanings. Note that the permittivity is a piecewise constant function of spatial variables, and in the grating region it is a periodic function of x only:

$$\varepsilon(x) = \begin{cases} \varepsilon_a, & \text{if } |x| \le d_1/2, \quad 0 \le z \le h; \\ \varepsilon_b, & \text{if } d_1/2 < |x| \le d/2, \quad 0 \le z \le h. \end{cases} \tag{13.1}$$

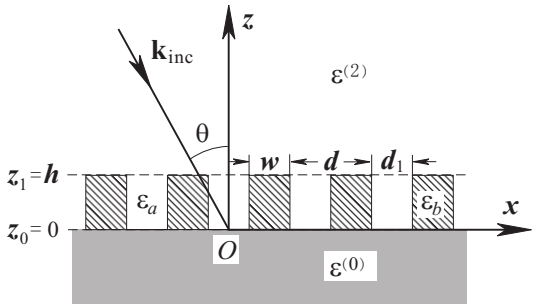


Fig. 13.1. Definition and notation of a rectangular grating problem.

The incident plane wave is characterized by its wavelength λ, wave vector $\mathbf{k}_{\text{inc}}$, and polarization vector $\hat{\mathbf{a}}$, where

$$\mathbf{k}_{\text{inc}} = k_0\sqrt{\varepsilon^{(2)}\mu^{(2)}}\,(\hat{\mathbf{x}}\sin\theta - \hat{\mathbf{z}}\cos\theta), \tag{13.2}$$

with $k_0 = 2\pi/\lambda$ being the vacuum wave vector. So, the plane of incidence coincides with the principal plane of the grating.

Since the grating structure is *y* invariant and the incident wave is independent of *y*, the electric and magnetic field vectors **E** and **H** are also independent of *y*. In this section we consider the TM polarization (transverse magnetic: the magnetic field vector perpendicular to the principal plane). Because we formally keep magnetic permeability μ in all formulas derived in this chapter, results for TE polarization can be easily obtained from the TM results by using the symmetry of Maxwell equations with respect to electric and magnetic quantities. The assumed harmonic time dependence in this chapter is exp(−*i*ωt).

13.2.2.2 Construction of the total fields

For the TM polarization problem it is convenient to work with the *y* component of the magnetic vector H_y. It has been shown in Chapter 2 that the total fields in the upper and lower semi-infinite regions can be readily written as Rayleigh expansions:

$$H_y(x,z) = \exp(\mathrm{i}\alpha_0 x - \mathrm{i}\gamma_0^{(2)} z) + \sum_{m=-\infty}^{+\infty} R_m^{(\mathrm{h})} \exp(\mathrm{i}\alpha_m x + \mathrm{i}\gamma_m^{(2)} z), \qquad z \ge h; \tag{13.3a}$$

$$H_y(x,z) = \sum_{m=-\infty}^{+\infty} T_m^{(\mathrm{h})} \exp(\mathrm{i}\alpha_m x - \mathrm{i}\gamma_m^{(0)} z), \qquad z \le 0. \tag{13.3b}$$

In the above equations

$$\alpha_m = \alpha_0 + mK, \qquad \alpha_0 = k^{(2)}\sin\theta, \qquad K = 2\pi/d, \tag{13.4}$$

$$\gamma_m^{(p)} = \sqrt{k^{(p)2} - \alpha_m^2}, \qquad \mathrm{Re}[\gamma_m^{(p)}] + \mathrm{Im}[\gamma_m^{(p)}] > 0, \qquad p = 0, 2 \tag{13.5}$$

where $k^{(p)2} = k_0^2 n^{(p)2}$ with $n^{(p)} = \sqrt{\varepsilon^{(p)}\mu^{(p)}}$ being the refractive indices of the two media, and $R_m^{(\mathrm{h})}$ and $T_m^{(\mathrm{h})}$ are the unknown Rayleigh coefficients. The first term in (13.3a) represents the incident plane wave of unit amplitude. The + and – signs in front of the second terms in the exponential functions in (13.3) have been carefully chosen in accordance with the inequality in (13.5) to ensure that the radiation condition is satisfied.

Next we find the total magnetic field in the periodic layer. To do so we need first to derive the eigenvalue equation that determines the modes of the periodic layer. For time-harmonic electromagnetic fields two of the Maxwell equations containing the curl operator are

$$\nabla\times\mathbf{E} = \mathrm{i}\,k_0\mu\mathbf{H}, \tag{13.6a}$$

$$\nabla\times\mathbf{H} = -\mathrm{i}\,k_0\varepsilon\,\mathbf{E}. \tag{13.6b}$$

For *y* independent TM polarized fields these equations become

$$\partial_z E_x - \partial_x E_z = \mathrm{i}\,k_0\mu H_y, \tag{13.7a}$$

$$\partial_x H_y = -\mathrm{i}\,k_0\varepsilon E_z, \tag{13.7b}$$

$$\partial_z H_y = \mathrm{i}\,k_0\varepsilon E_x. \tag{13.7c}$$

Using the last two equations to eliminate E_x and E_z and keeping in mind that ε is a function of only *x*, one can easily find

$$k_0^2 \varepsilon \mu H_y(x,z) + \varepsilon \partial_x \frac{1}{\varepsilon} \partial_x H_y(x,z) = -\partial_z^2 H_y(x,z) \,. \tag{13.8}$$

Since the coefficients of this differential equation do not depend on z, applying the standard procedure of separation of variables shows that the equation has a solution of the form $H_y(x,z) = H_y(x)\exp(i\gamma z)$, where γ is a constant. Substituting this into (13.8) yields

$$k_0^2 \varepsilon \mu H_y + \varepsilon \frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{\varepsilon} \frac{\mathrm{d}}{\mathrm{d}x} H_y = \gamma^2 H_y \,. \tag{13.9}$$

Henceforth H_y without an explicit variable dependence denotes the function $H_y(x)$. The ordinary differential equation (13.9) together with the associated boundary conditions defines an eigenvalue problem with γ being the eigenvalue. Here the boundary conditions are of two kinds. The first kind is the electromagnetic boundary conditions at the permittivity jump discontinuities. Obviously, being a tangential component to the medium interface H_y is continuous, and it follows from (13.7b) $(1/\varepsilon)\partial_x H_y$ must also be continuous:

$$H_y(\pm d_1/2 \pm 0) = H_y(\pm d_1/2 \mp 0), \quad \frac{1}{\varepsilon_b} \frac{\mathrm{d}H_y}{\mathrm{d}x}(\pm d_1/2 \pm 0) = \frac{1}{\varepsilon_a} \frac{\mathrm{d}H_y}{\mathrm{d}x}(\pm d_1/2 \mp 0) \,. \tag{13.10}$$

By the way, if (13.9) is understood in the sense of distribution, (13.10) is already contained in it. The second kind of boundary condition to be imposed on (13.9) is the pseudo-periodicity condition. Since $\varepsilon(x+d) = \varepsilon(x)$ appears in the coefficients of (13.9), Floquet theorem requires that $H_y(x)$ be pseudo-periodic

$$H_y(x+d) = \exp(\mathrm{i}\alpha_0 d) H_y(x) \,, \tag{13.11}$$

where α_0 is the pseudo-periodicity constant determined by the incident plane wave.

A solution of the eigenvalue problem defined by (13.9-11) is an eigenvalue and eigenfunction pair $\{\gamma, H_y(x;\gamma)\}$. An eigen-function $H_y(x;\gamma)$ is in physical term a mode of the lamellar grating structure. For mode $H_y(x;\gamma)$ the law of motion is (13.9) and the internal boundary conditions are (13.10) and (13.11). Effectively the problem is that of finding modes of a periodic waveguide. Because $H_y(x,z) = H_y(x;\gamma)\exp(i\gamma z)$ the x dependence of the modal field is independent of z. In searching for the eigen-solutions the z invariance of ε is used, as if the grating layer is infinitely thick. The finiteness of the grating layer is not recalled until the external boundary conditions are matched between different regions in the next step.

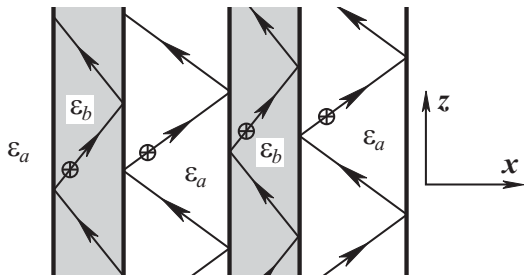


Fig. 13.2. Geometric optics picture of a mode of a periodic waveguide. The arrows represent optical rays and $\oplus$ represent the direction of the electric vector in TE polarization or that of the magnetic vector in TM polarization.

For a piecewise homogeneous periodic region the modal function has a simple geometric optics interpretation. In each medium the function is composed of two plane waves whose wave vectors share the same z component γ and have x components of equal length but opposite signs, as shown in Fig. 13.2. The zigzag paths can be understood as optical rays. In differ-

ent media the zigzag paths make different angles with respect to the x axis, but the z components of their associated wave vectors are the same. Since different modes have different eigenvalues γ the angles of zigzag paths for different modes within a given medium are also different. The phases of the ray paths at the same ordinate z and two abscissas differing by d are related by Floquet theorem. The planes of all zigzag paths are perpendicular to the medium interfaces. The periodic region allows two characteristic polarizations, one with its magnetic field vector perpendicular to the zigzag plane, which is the present case, and the other with its electric field vector perpendicular to the zigzag plane, which corresponds to TE polarization incidence. The zigzag path angles for TE polarization are in general different from those for TM polarization.

The eigenvalue spectrum of (13.9) is composed of a discrete set of numbers. Since γ appears in (13.9) as γ^2, if γ is an eigenvalue, so is $-\gamma$, and $\pm\gamma$ share the same eigen-function $H_y(x)$. In actual manipulation γ is chosen as one of the two square roots of γ^2. We only need to label half of the eigenvalues using index $q = 1, 2, \ldots$; the second half is referenced by adding a negative sign in front of the first half. When we limit our domain of discussion within the linear electromagnetic problems a linear superposition of modal fields is also a solution of Maxwell equations. A general solution in the grating layer satisfying all internal boundary conditions is given by a superposition of all possible modes

$$H_y(x,z) = \sum_{q=1}^{\infty} [\, u_q \exp(\mathrm{i}\gamma_q z) + d_q \exp(-\mathrm{i}\gamma_q z)\,] H_{y,q}(x), \tag{13.12}$$

where u_q and d_q are constants, and a comma is used in the subscript for $H_{y,q}(x)$ for a reason to become clear soon. When $H_{y,q}(x)$ are properly normalized, u_q and d_q are the modal amplitudes evaluated at $z = 0$. Since (13.12) is valid for z being finite, no radiation condition is to be imposed and both terms with the + and – signs in the exponential functions should be kept. In this sense it is not important how γ_q is chosen from the two square roots. However, to attach an unambiguous physical meaning to each term and to facilitate later numerical treatment, eigenvalue partition must be made. As with the $\gamma_m^{(l)}$ appearing in Rayleigh expansions, we require

$$\gamma_q = \sqrt{\gamma_q^2}\,, \qquad \mathrm{Re}[\gamma_q] + \mathrm{Im}[\gamma_q] > 0. \tag{13.13}$$

Then, the first term in (13.12) represents a mode that propagates or decays in the upward direction, and the second term represents a mode in the downward direction, justifying the choice of letters for the two superposition constants.

So far we have not discussed how to obtain the eigen-functions $H_{y,q}(x)$. In the Fourier modal method they are obtained by solving (13.9) in the discrete Fourier space. For this purpose $H_y(x)$ is expanded into Floquet-Fourier series

$$H_y(x) = \sum_{m=-\infty}^{\infty} H_{ym} \exp(\mathrm{i}\alpha_m x), \tag{13.14}$$

where α_m is defined in (13.4) and H_{ym} are the Fourier coefficients (for convenience we will indistinguishably refer to both Fourier and Floquet-Fourier coefficients as Fourier coefficients). Note that to distinguish from the qth eigen-function $H_{y,q}(x)$, the subscript in H_{ym} does not contain a comma. In this chapter when a roman letter subscript in italics such as m and n is attached to a symbol representing a function of spatial variables, the composite symbol denotes the Fourier coefficients of the function.

Equation (13.9) also contains $\varepsilon(x)$ that too has to be transformed into Fourier space. It is tempting to expand $\varepsilon(x)$ into a Fourier series and substitute it, together with (13.14), into (13.9) to derive the desired result. This is how it was done before 1996; however, numerically the resulting solution converged very slowly when medium a or b was a highly conducting metal.

It was later found that only when the Fourier coefficients were used in a special way the Fourier modal method could converge well for metallic rectangular gratings [13.10, 13.11]. This finding led to establishment of a theory called Fourier factorization [13.8], which not only explained the observed improvement of convergence in [13.10, 13.11] but also enabled a series of other improvements in grating theory. However, it would take us afield to cover this topic at length in this chapter, so I am content with giving only key results here. The interested reader can find more information in [13.5, 13.8], and Chapter 7 of this book.

Given a function $h(x)$ as the product of two periodic, piecewise continuous functions $g(x)$ and $f(x)$, all three functions having the same period, to compute the Fourier coefficients of $h(x)$ in terms of the Fourier coefficients of $g(x)$ and $f(x)$ one should follow the Fourier factorization rules:

(1) If $g(x)$ is discontinuous and $f(x)$ is continuous, then

$$h_{n,\Omega} = \sum_{m\in\Omega} g_{n-m} f_m, \qquad n \in \Omega. \tag{13.15}$$

(2) If both $g(x)$ and $f(x)$ are discontinuous but $g(x)f(x)$ is continuous, then

$$h_{n,\Omega} = \sum_{m\in\Omega} \left[\!\!\left[\frac{1}{g} \right]\!\!\right]^{-1}_{\Omega,nm} f_m, \qquad n \in \Omega. \tag{13.16}$$

(3) If $g(x)$ and $f(x)$ meet neither of the above two conditions, do not use (13.15) or (13.16) directly. Instead, rearrange $g(x)f(x)$ as a sum of several terms so that each term meets one of the two conditions, and then follow (1) and (2).

In the above Ω denotes a set of integers: $m_1 \le m \le m_2$ for some fixed m_1 and m_2. $[\![a]\!]$ Denotes the square Toeplitz matrix generated by the Fourier coefficients of function $a(x)$ such that $[\![a]\!]_{mn} = a_{m-n}$, and $[\![a]\!]^{-1}$ means the matrix inverse of $[\![a]\!]$. The subscript Ω on the right-hand side of (13.16) is used to indicate the size of the matrix and that on the left-hand side of both equations are to indicate that these Fourier coefficients depend on Ω. When there is no danger of confusion, Ω can be omitted.

The three types of products in (1), (2), and (3) are referred to as Type 1, Type 2, and Type 3 products, respectively. From a mathematical point of view (13.15) and (13.16) give two multiplication rules for multiplying two Fourier series in a truncated Fourier space delimited by Ω. The one in (13.15) is called Laurent's rule and that in (13.16) is called the inverse rule. Although the factorization rules are stated for $h(x)$, $f(x)$, and $g(x)$ all being periodic functions, they are obviously valid when $h(x)$ and $f(x)$ are pseudo-periodic because we can multiply $h(x) = g(x)\, f(x)$ by $\exp(-i\alpha_0 x)$.

The essence of following the Fourier factorization rules is to preserve continuity nature of the product function. While the pseudo-periodicity condition is automatically satisfied once $H_{ay}(x)$ is expanded into Floquet-Fourier series as in (13.14), the internal electromagnetic boundary conditions (13.10) have to be expressly enforced. Equation (13.9) contains three products of periodic and pseudo-periodic functions: εH_y, $(1/\varepsilon)H_y'$, and $\varepsilon\,[(1/\varepsilon)H_y']'$, where the prime denotes derivative with respect to x. They are of Types 1, 2, and 3, respectively. The presence of the Type 3 product prevents a direct Fourier factorization of (13.9). The difficulty can be avoided by rewriting (13.9) as

$$\varepsilon\left(k_0^2 \mu H_y + \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\varepsilon}\frac{\mathrm{d}}{\mathrm{d}x} H_y \right) = \gamma^2 H_y\,. \tag{13.17}$$

This new equation contains only Type 2 products. Before making the Fourier transform let us note that the action of the differential operator d/dx on a pseudo-periodic series is to multiply its nth term by $i\alpha_n$ for all n. Then the image of (13.9) in discrete Fourier space is

$$\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\left(k_0^2\mu\mathbf{I}-\boldsymbol{\alpha}\left[\!\left[\varepsilon\right]\!\right]^{-1}\boldsymbol{\alpha}\right)H_y=\gamma^2 H_y\,, \tag{13.18}$$

where $\mathbf{I}$ is the identity matrix, $\boldsymbol{\alpha}$ is a diagonal matrix with α_n being its diagonal elements, and without possibility of confusion we have used H_y to denote the column vector formed by H_{ym}, the Fourier coefficients of $H_y(x)$. Note that equation (13.18) contains all three ingredients of our boundary value problem in one equation.

Equation (13.18) can also be derived by first transforming (13.7) into Fourier space and then eliminating the column vectors formed by the Fourier coefficients of $E_x(x)$ and $E_z(x)$. We will see in later sections, it is always easier to transform the original Maxwell equations into Fourier space when products of functions are simple and then to carry out the necessary manipulations in Fourier space, than to defer the transformation until a later or the final stage when the products of functions involved in a wave equation have become more complicated.

Equation (13.18) is written in the standard form of a matrix eigenvalue problem. It can be solved numerically by using an eigenvalue-eigenvector solver available in many computer software packages. When the matrix in (13.18) is truncated to $N \times N$, where $N = m_2 - m_1 + 1$ is the number of elements in set Ω, the number of eigen-solutions is N, but after taking square roots of γ^2 there are $2N$ eigen-solutions to the underlying physical problem. Criterion (13.13) may be used to make the eigenvalue spectrum partition. In comparison with (13.12) the total magnetic field is now written as a Floquet-Fourier series

$$H_y(x,z)=\sum_{m=m_1}^{m_2}\exp(\mathrm{i}\alpha_m x)\sum_{q=1}^{N}H_{ymq}\left[u_q\exp(\mathrm{i}\gamma_q z)+d_q\exp(-\mathrm{i}\gamma_q z)\right], \tag{13.19}$$

where H_{ymq} are the Fourier coefficients of $H_{y,q}(x)$. So far we have derived the general expressions of the total magnetic field in all three spatial regions. These expressions contain the unknown modal amplitudes u_q and d_q and the unknown Rayleigh coefficients (that can be considered as special modal amplitudes).

Before moving on to solve for the unknown coefficients we comment on the practical issue of truncation. The truncation number N determines the accuracy of the numerical results; the larger N is the more accurate the results are but the more computation time and computer memory are consumed. N should be chosen large enough to include all propagating orders and sufficiently many evanescent orders on both side of the propagating orders. How large is sufficient depends on the specific problem. It is difficult to give a general criterion. Certainly a metallic grating requires a larger N than does a dielectric grating, and a grating of near zero or near 100% duty cycle requires a larger N than does a grating of 50% duty cycle. The best way to be sure is to run a few convergence tests. The minimum N beyond which the numerical results of quantities of interest stabilize is a good truncation number to use. It is a common practice that the truncation interval $[m_1,m_2]$ is chosen as $[-M,M]$ for some $M > 0$; however, it is better to set $m_1 = -[\alpha_0/K] - [N/2]$, where $[\cdot]$ means the integral part of. In this way good truncation coverage is achieved even when the angle of incidence is near grazing.

13.2.2.3 Matching of the external boundary conditions

In the Fourier modal method the matching of boundary conditions between different regions is realized by matching the Fourier coefficients of the tangential components of the total fields. In the preceding subsection the total magnetic fields in the three regions have been written in Floquet-Fourier series. The other tangential component, the x component of the electric field,

is yet to be found and written in this form. Equation (13.7c) provides the link between E_x and H_y:

$$E_x = \frac{1}{\mathrm{i}k_0\varepsilon}\partial_z H_y \,. \tag{13.20}$$

In regions 0 and 2, ε is a constant, so it follows easily from (13.3) that

$$E_x(x,z) = \frac{1}{k_0\varepsilon^{(2)}}\left[-\gamma_0^{(2)}\exp(\mathrm{i}\alpha_0 x - \mathrm{i}\gamma_0^{(2)} z) + \sum_{m=-\infty}^{+\infty}\gamma_m^{(2)} R_m^{(\mathrm{h})}\exp(\mathrm{i}\alpha_m x + \mathrm{i}\gamma_m^{(2)} z)\right], \quad z \geq h; \tag{13.21a}$$

$$E_x(x,z) = -\frac{1}{k_0\varepsilon^{(0)}}\sum_{m=-\infty}^{+\infty}\gamma_m^{(0)} T_m^{(\mathrm{h})}\exp(\mathrm{i}\alpha_m x - \mathrm{i}\gamma_m^{(0)} z), \quad z \leq 0. \tag{13.21b}$$

In the periodic layer, the Floquet-Fourier series of E_x in terms of H_y is obtained by transforming the right-hand side of (13.20) into Fourier space using Laurent's rule:

$$E_x(x,z) = \frac{1}{k_0}\sum_{m=m_1}^{m_2}\exp(\mathrm{i}\alpha_m x)\sum_{n=m_1}^{m_2}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]_{mn}\sum_{q=1}^{N}\gamma_q H_{ynq}\left[u_q\exp(\mathrm{i}\gamma_q z) - d_q\exp(-\mathrm{i}\gamma_q z)\right]. \tag{13.22}$$

Equations (13.3), (13.19), (13.21), and (13.22) provide the Fourier coefficients of all vector field components needed to match boundary conditions.

To facilitate subsequent development it is convenient to define two column vectors. The first is the vector of modal amplitudes:

$$F = \begin{pmatrix}\tilde{u}_q \\ \tilde{d}_q\end{pmatrix}, \tag{13.23}$$

where $\tilde{u}_q$ and $\tilde{d}_q$ are the qth upward and downward modal amplitudes evaluated at the lower and upper boundaries of the grating layer, respectively:

$$\tilde{u}_q = u_q\exp(\mathrm{i}\gamma_q z_0), \qquad \tilde{d}_q = d_q\exp(-\mathrm{i}\gamma_q z_1). \tag{13.24}$$

In (13.23) a generic vector element is used to represent the whole sub-vector. So, if the number of modes to be used in numerical computation is N, then $\tilde{u}_q$ and $\tilde{d}_q$ each represents an $N \times 1$ column vector with $1 \leq q \leq N$. (13.23) is written for the periodic layer. In the upper semi-infinite space $\tilde{u}_q$ and $\tilde{d}_q$ are to be replaced by $\tilde{R}_m^{(\mathrm{h})}$ and $\tilde{\delta}_{mn}$, and in the lower semi-infinite space by 0 and $\tilde{T}_m^{(\mathrm{h})}$, respectively, where

$$\tilde{R}_m^{(\mathrm{h})} = R_m^{(\mathrm{h})}\exp(\mathrm{i}\gamma_m^{(2)} z_1), \quad \tilde{T}_m^{(\mathrm{h})} = T_m^{(\mathrm{h})}\exp(-\mathrm{i}\gamma_m^{(0)} z_0), \quad \tilde{\delta}_{mn} = \delta_{mn}\exp(-\mathrm{i}\gamma_m^{(2)} z_2), \tag{13.25}$$

and $\delta_{mn} = 0$ for $m \neq n$ and $\delta_{mn} = 1$ for $m = n$. The second vector to define is the vector of Fourier coefficients of the total fields

$$\mathscr{F}(z) = \begin{pmatrix}H_{ym}(z) \\ E_{xm}(z)\end{pmatrix}, \tag{13.26}$$

where H_{ym} and E_{xm} are also short-hand notations of $N \times 1$ column vectors. In terms of these column vectors the Fourier coefficients of the total fields in the three spatial regions are

$$\mathscr{F}(z) = \mathbf{W}^{(2)}\phi^{(2)}(z)\begin{pmatrix}\tilde{R}_m^{(\mathrm{h})}\\ \tilde{\delta}_{m0}\end{pmatrix}, \qquad z \ge z_1; \tag{13.27a}$$

$$\mathscr{F}(z) = \mathbf{W}^{(1)}\,\phi^{(1)}(z)\begin{pmatrix}\tilde{u}_q\\ \tilde{d}_q\end{pmatrix}, \qquad z_0 \le z \le z_1; \tag{13.27b}$$

$$\mathscr{F}(z) = \mathbf{W}^{(0)}\phi^{(0)}(z)\begin{pmatrix}0\\ \tilde{T}_m^{(\mathrm{h})}\end{pmatrix}, \qquad z \le z_0, \tag{13.27c}$$

where

$$\phi^{(p)}(z) = \begin{pmatrix}\exp[\mathrm{i}\,\gamma_m^{(p)}(z - z_{p-1})] & 0\\ 0 & \exp[\mathrm{i}\,\gamma_m^{(p)}(z_p - z)]\end{pmatrix}, \qquad p = 0, 2, \tag{13.28a}$$

$$\phi^{(1)}(z) = \begin{pmatrix}\exp[\mathrm{i}\,\gamma_q\,(z - z_0)] & 0\\ 0 & \exp[\mathrm{i}\,\gamma_q\,(z_1 - z)]\end{pmatrix}, \tag{13.28b}$$

$$\mathbf{W}^{(p)} = \begin{pmatrix}\delta_{mn} & \delta_{mn}\\ \dfrac{\gamma_m^{(p)}}{k_0\varepsilon^{(p)}}\delta_{mn} & -\dfrac{\gamma_m^{(p)}}{k_0\varepsilon^{(p)}}\delta_{mn}\end{pmatrix}, \qquad p = 0, 2, \tag{13.29a}$$

$$\mathbf{W}^{(1)} = \begin{pmatrix}H_{ymq} & H_{ymq}\\ \dfrac{1}{k_0}\displaystyle\sum_{n=m_1}^{m_2}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]_{mn} H_{ynq}\gamma_q & -\dfrac{1}{k_0}\displaystyle\sum_{n=m_1}^{m_2}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]_{mn} H_{ynq}\gamma_q\end{pmatrix}. \tag{13.29b}$$

In all of the above matrices a generic matrix element represents a matrix. For example, H_{ymq} stands for the square matrix and $\exp(i\gamma_q z)$ stands for the diagonal matrix. All matrices in (13.27-29) are of finite size, truncated from $m = m_1$ to $m_2 = N + m_1 - 1$, or from $q = 1$ to N. Note that all $\mathbf{W}$ matrices have a symmetric structure: If W_{ij}, where $i, j = 1, 2$, are used to denote the 2×2 block matrices, then $W_{12} = W_{11}$ and $W_{22} = -W_{21}$. This symmetry property can be exploited to save computation time [13.12]. With all of the above notations the matching of the boundary conditions at $z = z_0$ and $z = z_1$ can be written as $\mathscr{F}(z_1 + 0) = \mathscr{F}(z_1 - 0)$:

$$\mathbf{W}^{(2)}\phi^{(2)}(z_1)\begin{pmatrix}\tilde{R}_m^{(\mathrm{h})}\\ \tilde{\delta}_{m0}\end{pmatrix} = \mathbf{W}^{(1)}\,\phi^{(1)}(z_1)\begin{pmatrix}\tilde{u}_q\\ \tilde{d}_q\end{pmatrix}, \tag{13.30a}$$

and $\mathscr{F}(z_0 + 0) = \mathscr{F}(z_0 - 0)$:

$$\mathbf{W}^{(1)}\,\phi^{(1)}(z_0)\begin{pmatrix}\tilde{u}_q\\ \tilde{d}_q\end{pmatrix} = \mathbf{W}^{(0)}\phi^{(0)}(z_0)\begin{pmatrix}0\\ \tilde{T}_m^{(\mathrm{h})}\end{pmatrix}. \tag{13.30b}$$

13.2.2.4 Solution of the boundary matching equations

Equations (13.30a) and (13.30b) has this general form

$$\mathbf{W}^{(p+1)}\begin{pmatrix}1 & 0\\ 0 & \exp(\mathrm{i}\gamma_q^{(p+1)}h_{p+1})\end{pmatrix}\begin{pmatrix}\tilde{u}_q^{(p+1)}\\ \tilde{d}_q^{(p+1)}\end{pmatrix}=\mathbf{W}^{(p)}\begin{pmatrix}\exp(\mathrm{i}\gamma_q^{(p)}h_p) & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tilde{u}_q^{(p)}\\ \tilde{d}_q^{(p)}\end{pmatrix}, \tag{13.31}$$

with

$$\tilde{u}_q^{(p)}=u_q^{(p)}\exp(\mathrm{i}\gamma_q^{(p)}z_{p-1}),\qquad \tilde{d}_q^{(p)}=d_q^{(p)}\exp(-\mathrm{i}\gamma_q^{(p)}z_p), \tag{13.32}$$

provided that $h_p = z_p - z_{p-1}$ and $(\tilde{u}_q^{(p)}, \tilde{d}_q^{(p)})^{\mathrm{T}}$ is understood as the column vector composed of the Rayleigh amplitudes when p refers to a semi-infinite medium. This type of boundary matching equation is not unique in the Fourier modal method. It is shared by many other grating methods although the compositions of their **W** matrices are different. When the boundary matching equations have been derived the grating problem is nearly solved. The remaining steps, which are more or less the same for all grating methods, are to solve the resulting linear system of equations for the unknown Rayleigh amplitudes $\tilde{R}_m^{(\mathrm{h})}$ and $\tilde{T}_m^{(\mathrm{h})}$, and possibly also the internal modal amplitudes $\tilde{u}_q$ and $\tilde{d}_q$. From these amplitudes the diffraction efficiencies and parameters of polarization states can be easily obtained.

In principle many algorithms can be used to solve the boundary matching equations. Because of the appearance of the exponential functions in (13.31), some of the solution methods are numerically stable and some are not. Among the numerically stable algorithms the best method is the S matrix propagation algorithm [13.12-14]. It has the advantages of having the most clear physical meaning and the best computation efficiency. Many variants of the S matrix algorithm exist; one of them is described in Section 13.5 for the general multilayered grating problem.

13.2.2.5 Final solution of the grating problem

The diffraction efficiency of a propagating order is defined as the ratio of the z component of the Poynting vector of that order to the z component of the Poynting vector of the incident plane wave. After $R_m^{(\mathrm{h})}$ and $T_m^{(\mathrm{h})}$ have been obtained the TM diffraction efficiencies can be easily found (cf. Chap. 2):

$$\eta_{m,\mathrm{TM}}^{(2)}=\frac{\gamma_m^{(2)}}{\gamma_0^{(2)}}|R_m^{(\mathrm{h})}|^2,\qquad m\in U^{(2)}, \tag{13.33a}$$

$$\eta_{m,\mathrm{TM}}^{(0)}=\frac{\varepsilon^{(2)}\gamma_m^{(0)}}{\varepsilon^{(0)}\gamma_0^{(2)}}|T_m^{(\mathrm{h})}|^2,\qquad m\in U^{(0)}, \tag{13.33b}$$

where $\eta_{m,\mathrm{TM}}^{(2)}$ and $\eta_{m,\mathrm{TM}}^{(0)}$ are the mth-order diffraction efficiencies in the reflection and transmission sides of the grating, respectively, and $U^{(p)}$ is the set of propagating order numbers in medium p, $p = 0, 2$.

In this section we have dealt with only the TM polarization. The results for TE polarization can be easily obtained by formally making exchanges $\mathbf{E} \leftrightarrow \mathbf{H}$ and $\varepsilon \leftrightarrow -\mu$ and changing the superscripts of the Rayleigh amplitudes from (h) to (e) in all formulas derived in this section. In particular, the TE diffraction efficiencies are

$$\eta_{m,\mathrm{TE}}^{(2)}=\frac{\gamma_m^{(2)}}{\gamma_0^{(2)}}|R_m^{(\mathrm{e})}|^2,\qquad m\in U^{(2)}, \tag{13.34a}$$

$$\eta_{m,\mathrm{TE}}^{(0)}=\frac{\mu^{(2)}\gamma_m^{(0)}}{\mu^{(0)}\gamma_0^{(2)}}|T_m^{(\mathrm{e})}|^2,\qquad m\in U^{(0)}. \tag{13.34b}$$

Note that because the definitions of $\gamma_m^{(p)}$ are independent of polarization, the diffraction angles and the number of propagating orders are also independent of polarization.

13.2.3 Formulation in an oblique Cartesian coordinate system

13.2.3.1 Description of the problem of slanted gratings

The rectangular grating shown in Fig. 13.1 has the mirror reflection symmetry with respect to the *yz* plane. In many applications it is desirable to break this symmetry. A surface relief grating with its two sidewalls of a ridge parallel to each other but tilted with respect to the *yz* plane is called a slanted grating. As a typical example a parallelogram grating is shown in Fig. 13.3. All previously defined symbols have the same meanings as in Fig. 13.1. The angle ζ is called the slant angle of the grating (in the figure $\zeta > 0$). *D* is the period of the slanted slabs or fringes. The groove depth measured along the slant direction is $\tilde{h} = h/\cos\zeta$. For this type of gratings it is obvious that the rectangular staircase approximation as illustrated in the left part of the figure is very inefficient.

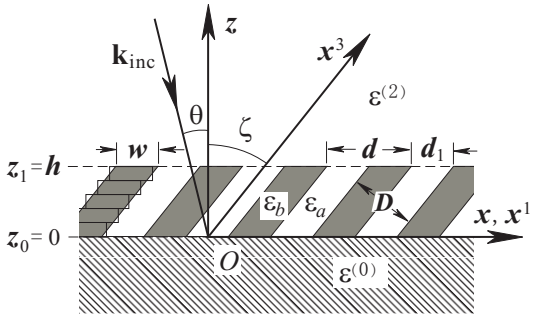


Fig. 13.3. Definition and notation of a slanted grating problem.

So far the theoretical treatments of slanted gratings have been made almost exclusively in rectangular Cartesian coordinate systems. Some authors use only one rectangular system like the one in Fig. 13.3, and others use a second rectangular system that is aligned with the surface and the normal of groove sidewalls. In either case the resulting mathematical treatments are difficult, if not awkward. However, there is nothing sacred about the rectangular coordinate system. For a parallelogram grating it is the most natural to use the oblique Cartesian coordinate system as shown in Fig. 13.3, in which every piece of flat surfaces of the grating coincides with a coordinate surface. It will be demonstrated that the Fourier modal method can be formulated the most efficiently in the matched oblique coordinate system. A small price to pay is to be familiar with properties of the uncommon oblique coordinate system.

13.2.3.2 Oblique Cartesian coordinate system

In Fig. 13.3 the rectangular coordinate system *Oxyz* is set up in the same way as in Fig. 13.1, with the *y* axis parallel to the groove direction and the *z* axis perpendicular to the grating plane. The oblique coordinate system $Ox^1x^2x^3$ is set up such that the x^1 and x^2 axes are parallel to the *x* and *y* axes, respectively, and the x^3 axis is in the *xz* plane and forms an angle ζ with respect to the *z* axis. The coordinate transformation from *Oxyz* to $Ox^1x^2x^3$ is

$$x^1 = x - z\tan\zeta, \qquad x^2 = y, \qquad x^3 = z/\cos\zeta, \tag{13.35a}$$

and the inverse transformation is

$$x = x^1 + x^3\sin\zeta, \qquad y = x^2, \qquad z = x^3\cos\zeta. \tag{13.35b}$$

For the oblique coordinate system we define three basis vectors $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ of unit length that are in the positive directions of the x^1, x^2, and x^3 axes, respectively. From Fig. 13.3 it follows that in terms of basis vectors of *Oxyz*,

$$\mathbf{b}_1 = \hat{\mathbf{x}}, \qquad \mathbf{b}_2 = \hat{\mathbf{y}}, \qquad \mathbf{b}_3 = \hat{\mathbf{x}} \sin\zeta + \hat{\mathbf{z}} \cos\zeta . \tag{13.36}$$

We define another set of basis vectors, not necessarily of unit length by

$$\mathbf{b}^1 = \frac{\mathbf{b}_2 \times \mathbf{b}_3}{\sqrt{g}}, \qquad \mathbf{b}^2 = \frac{\mathbf{b}_3 \times \mathbf{b}_1}{\sqrt{g}}, \qquad \mathbf{b}^3 = \frac{\mathbf{b}_1 \times \mathbf{b}_2}{\sqrt{g}}, \tag{13.37}$$

where $\sqrt{g} = \mathbf{b}_1 \cdot \mathbf{b}_2 \times \mathbf{b}_3 = \cos\zeta$. From (13.36) it follows that

$$\mathbf{b}^1 = \hat{\mathbf{x}} - \hat{\mathbf{z}} \tan\zeta , \qquad \mathbf{b}^2 = \hat{\mathbf{y}}, \qquad \mathbf{b}^3 = \hat{\mathbf{z}} / \cos\zeta . \tag{13.38}$$

The basis vectors $\mathbf{b}^1$, $\mathbf{b}^2$, and $\mathbf{b}^3$ are called reciprocal space basis vectors or contravariant basis vectors, and $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ are called the real space basis vectors or covariant basis vectors. The two sets are mutually orthonormal: $\mathbf{b}_\sigma \cdot \mathbf{b}^\rho = \delta_\sigma^\rho$, where $\delta_\sigma^\rho = 1$ if $\sigma = \rho$ and $\delta_\sigma^\rho = 0$ if $\sigma \neq \rho$. Figure 13.4 illustrates the relationship between $\mathbf{b}_\sigma$ and $\mathbf{b}^\rho$ in the x^1x^3 plane. We define covariant and contravariant metric tensors as

$$g_{\rho\sigma} = \mathbf{b}_\rho \cdot \mathbf{b}_\sigma , \qquad g^{\rho\sigma} = \mathbf{b}^\rho \cdot \mathbf{b}^\sigma , \tag{13.39}$$

respectively. In this chapter we use lowercase Greek letters ρ, σ, and τ to label components of vectors and tensors, lowercase roman letters *i*, *j*, *m*, and *n* to label Fourier coefficients of field components and permittivity and permeability tensors, and letter *q* to label eigensolutions.

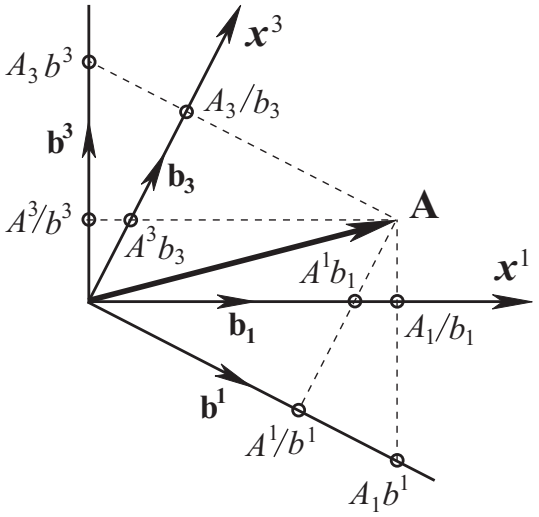


Fig. 13.4. Relationships between basis vectors and geometric interpretations of covariant and contravariant components of a vector.

An arbitrary vector can be expressed using covariant or contravariant basis vectors:

$$\mathbf{A} = A_\sigma \mathbf{b}^\sigma = A^\sigma \mathbf{b}_\sigma , \tag{13.40}$$

where summation from 1 to 3 with respect to a pair of subscript and superscript of the same symbol is implied and A_σ and A^σ are called the covariant and contravariant components of vector $\mathbf{A}$, respectively. The metric tensors provide the links between A_σ and A^σ:

$$A_\rho = g_{\rho\sigma} A^\sigma , \qquad A^\rho = g^{\rho\sigma} A_\sigma . \tag{13.41}$$

The geometrical meanings of these vector components can be summarized as follows. A_ρ is the perpendicular projection of **A** on the $\mathbf{b}_\rho$ direction in unit of $1/b_\rho$ and the parallel projection of **A** on the $\mathbf{b}^\rho$ direction in unit of b^ρ; A^ρ is the parallel projection of **A** on the $\mathbf{b}_\rho$ direction in unit of b_ρ and the perpendicular projection of **A** on the $\mathbf{b}^\rho$ direction in unit of $1/b^\rho$. The physical significance of these geometrical properties lies in the fact that many physical laws are expressed in terms of vector components in the sense of perpendicular projections. For example, the electromagnetic boundary conditions are the continuities of the tangential components of the electric and magnetic field vectors and the normal components of the electric displacement and magnetic inductance vectors across a medium interface. Here tangential and normal components are understood as the perpendicular projections of the vectors. In solving an electromagnetic boundary value problem whenever possible it is best to set up the coordinate system so that the medium interfaces are coordinate surfaces. Then the covariant and the contravariant components are proportional to the tangential and normal components, respectively, to the medium interfaces.

13.2.3.3 Construction of the total fields

To facilitate the matching of boundary conditions in the present case the Rayleigh expansions should be written in the oblique coordinates. By following steps similar to those taken in Chap. 2 for deriving Rayleigh expansions in the rectangular coordinate system, one can find

$$H_2(x^1,x^3)=\exp(\mathrm{i}\alpha_0 x^1+\mathrm{i}\gamma_0^{(2)-}x^3)+\sum_{m=-\infty}^{+\infty}R_m^{(\mathrm{h})}\exp(\mathrm{i}\alpha_m x^1+\mathrm{i}\gamma_m^{(2)+}x^3),\quad x^3\ge\tilde{h};\tag{13.42a}$$

$$H_2(x^1,x^3)=\sum_{m=-\infty}^{+\infty}T_m^{(\mathrm{h})}\exp(\mathrm{i}\alpha_m x^1+\mathrm{i}\gamma_m^{(0)-}x^3),\qquad x^3\le 0,\tag{13.42b}$$

where

$$\begin{aligned}&\gamma_m^{(p)\pm}=\alpha_m\sin\zeta\pm\sqrt{k^{(p)2}-\alpha_m^2}\,\cos\zeta,\\&\mathrm{Re}[\sqrt{k^{(p)2}-\alpha_m^2}\,]+\mathrm{Im}[\sqrt{k^{(p)2}-\alpha_m^2}\,]>0,\end{aligned}\qquad p=0,2,\tag{13.43}$$

and α_m are still given by (13.4). Evidently here $\gamma_m^{(p)\pm}$ play the roles that $\pm\gamma_m^{(p)}$ play in (13.3). The square root selection criterion in (13.43) ensures that the terms in (13.42) have proper physical meanings.

To find the modal fields in the periodic layer we use the two curl equations of Maxwell written in the oblique Cartesian coordinate form:

$$\xi^{\rho\sigma\tau}\partial_\sigma E_\tau=\mathrm{i}k_0\mu\sqrt{g}\,H^\rho,\tag{13.44a}$$

$$\xi^{\rho\sigma\tau}\partial_\sigma H_\tau=-\mathrm{i}k_0\varepsilon\sqrt{g}\,E^\rho,\tag{13.44b}$$

where $\xi^{\rho\sigma\tau}=+1$ or -1 when $\{\rho,\sigma,\tau\}$ is an even or odd permutation of $\{1,2,3\}$, respectively, and $\xi^{\rho\sigma\tau}=0$ if any two of the indices are the same. Equation (44) is valid for any curvilinear coordinate systems. Its proof can be found in many standard textbooks on tensor analysis. Since we are still dealing with non-conical diffraction, in which the two fundamental polarization cases are decoupled, we can consider TM polarization only. It follows from (13.44)

$$\partial_3E_1-\partial_1E_3=\mathrm{i}k_0^*\mu H_2,\tag{13.45a}$$

$$\partial_1H_2=-\mathrm{i}k_0^*\varepsilon E^3,\tag{13.45b}$$

$$\partial_3H_2=\mathrm{i}k_0^*\varepsilon E^1,\tag{13.45c}$$

where $k_0^* = k_0\cos\zeta$ and we have used the fact that $H^2 = H_2$ for the particular coordinate system (13.35).

We now transform the equations in (13.45) into Fourier space while they are still in simple form. The modal field components are first expanded into Floquet-Fourier series, with an exponential x^3 dependence,

$$F(x^1, x^3) = \exp(\mathrm{i}\gamma x^3) \sum_{m=-\infty}^{\infty} F_m \exp(\mathrm{i}\alpha_m x^1), \tag{13.46}$$

where F stands for H_2, E_1, etc. The transformation of (13.45a) is straightforward, and that of (13.45c) requires using the inverse rule because E^1 is proportional to the component of **E** normal to the slanted sidewalls. The product εE^3 in (13.45b) is type 3 and requires some rearrangement. From (13.39)

$$E^1 = \frac{1}{\cos^2\zeta}(E_1 - \sin\zeta E_3), \tag{13.47a}$$

$$E^3 = E_3 - \sin\zeta E^1. \tag{13.47b}$$

Substitution of (13.47b) into (13.45b) allows Fourier factorization of the two terms using Laurent's rule and inverse rule separately. So, in Fourier space (13.45) becomes

$$\gamma E_1 - \boldsymbol{\alpha} E_3 = k_0^* \mu H_2, \tag{13.48a}$$

$$\boldsymbol{\alpha} H_2 = \frac{k_0}{\cos\zeta}\left[\sin\zeta \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} E_1 - \left(\left[\!\left[\varepsilon\right]\!\right]\cos^2\zeta + \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\sin^2\zeta\right)E_3\right], \tag{13.48b}$$

$$\gamma H_2 = \frac{k_0}{\cos\zeta}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}(E_1 - \sin\zeta E_3). \tag{13.48c}$$

In (13.48) H_2, E_1, and E_3 are understood as column vectors in Fourier space. Using (13.48b) to eliminate E_3 and after a little matrix algebraic manipulation we obtain

$$\begin{pmatrix} \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\mathbf{G}\boldsymbol{\alpha}\sin\zeta & k_0^*\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\mathbf{G}\left[\!\left[\varepsilon\right]\!\right] \\ \frac{k_0^*}{k_0^2}(k_0^2\mu\mathbf{I} - \boldsymbol{\alpha}\mathbf{G}\boldsymbol{\alpha}) & \sin\zeta\,\boldsymbol{\alpha}\mathbf{G}\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} \end{pmatrix}\begin{pmatrix} H_2 \\ E_1 \end{pmatrix} = \gamma\begin{pmatrix} H_2 \\ E_1 \end{pmatrix}, \tag{13.49}$$

where

$$\mathbf{G} = \left(\cos^2\zeta\,\left[\!\left[\varepsilon\right]\!\right] + \sin^2\zeta\,\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1}\right)^{-1}. \tag{13.50}$$

Equation (13.49) is the matrix eigenvalue problem for the modal field of the slanted grating region. This equation consolidates Maxwell equations, the pseudo-periodicity condition, and the internal boundary conditions at the medium interfaces into one. Each of its solution vectors gives the Fourier coefficients of both field components needed for matching the external boundary conditions. The geometric interpretation of modal field in Fig. 13.2 obviously applies here. The only needed change is to rotate the infinite periodic medium and the modal field in it around the y axis clockwise by angle ζ.

To numerically solve (13.49) we need first to truncate the matrix, which can be done in the same way as for the rectangular grating case. If the Floquet-Fourier series are truncated from m_1 to $m_2 = N + m_1 - 1$, then the coefficient matrix in (13.49) is $2N \times 2N$, which is twice

as large as the one for a rectangular grating. Note that γ no longer appears as γ^2 in the eigenvalue equation, so the $2N$ eigenvalues in general do not form ± pairs. This may appear to contradict the intuition that inside the periodic layer modes propagating in both positive and negative directions of the x^3 axis are possible. The paradox is resolved when one realizes that a set of modes are determined subject to a pseudo-periodicity condition. Although in the un-slanted grating case both upward and downward modes share the same pseudo-periodicity condition, in the slanted grating case modes having eigenvalues of the same magnitude but opposite signs generally do not share the same pseudo-periodicity condition. Nevertheless the pseudo-periodicity condition is uniquely determined by the incident angle θ and the true grating period d.

To ensure that the S matrix algorithm work properly it is only necessary to partition the eigenvalue spectrum so that eigenvalues with positive imaginary parts are in set σ^+ and eigenvalues with negative imaginary parts are in set σ^-. (Here the superscripted σ is not to be confused with the tensor index σ.) The real eigenvalues can be arbitrarily distributed into σ^+ and σ^- as long as the two sets each contain N elements. This arbitrary handling is permissible because in an internal layer both upward and downward propagating and decay modes are present, and due to the boundedness of propagating modes misidentification of their propagation directions does not lead to any numerical error (a rigorous partition according to the true physical nature of the modes is possible, but troublesome and hardly necessary). The eigenvalues in $\sigma^\pm$ may be labeled by q, with $q = 1, 2, \ldots, N$, and sorted in ascending order of absolute values of imaginary parts.

After (13.49) is solved the total fields in the periodic layer can be written as

$$\begin{pmatrix} H_2(x^1,x^3) \\ E_1(x^1,x^3) \end{pmatrix} = \sum_{m=m_1}^{m_2} \exp(\mathrm{i}\alpha_m x^1) \sum_{q=1}^{N} \begin{pmatrix} H_{2mq}^+ & H_{2mq}^- \\ E_{1mq}^+ & E_{1mq}^- \end{pmatrix} \begin{pmatrix} \exp[\mathrm{i}\gamma_q^+(x^3-x_0^3)] & 0 \\ 0 & \exp[\mathrm{i}\gamma_q^-(x^3-x_1^3)] \end{pmatrix} \begin{pmatrix} \tilde{u}_q \\ \tilde{d}_q \end{pmatrix}, \tag{13.51}$$

where $\tilde{u}_q = u_q \exp(\mathrm{i}\gamma_q^+ x_0^3)$ and $\tilde{d}_q = d_q \exp(\mathrm{i}\gamma_q^- x_1^3)$ are the phase-adjusted unknown modal amplitudes, $x_0^3 = 0$, and $x_1^3 = \tilde{h}$. For later convenience, we also define $x_{-1}^3 = x_0^3$, $x_2^3 = x_1^3$.

13.2.3.4 Remaining steps

Solution of (13.51) gives both H_2 and E_1 that are needed for matching boundary conditions from the grating region side. In the homogeneous regions we have yet to find E_1. Since $E_1 = E^1 + \sin\zeta\, E^3$, it follows from (13.42) and (13.45)

$$E_1(x^1,x^3) = \frac{\gamma_0^{(2)-} - \alpha_0 \sin\zeta}{k_0^* \varepsilon^{(2)}} \exp(\mathrm{i}\alpha_0 x^1 + \mathrm{i}\gamma_0^{(2)-} x^3) + \frac{1}{k_0^* \varepsilon^{(2)}} \times$$

$$\times \sum_{m=-\infty}^{+\infty} (\gamma_m^{(2)+} - \alpha_m \sin\zeta) R_m^{(\mathrm{h})} \exp(\mathrm{i}\alpha_m x^1 + \mathrm{i}\gamma_m^{(2)+} x^3), \qquad x^3 \ge \tilde{h}; \tag{13.52a}$$

$$E_1(x^1,x^3) = \frac{1}{k_0^* \varepsilon^{(0)}} \sum_{m=-\infty}^{+\infty} (\gamma_m^{(0)-} - \alpha_m \sin\zeta) T_m^{(\mathrm{h})} \exp(\mathrm{i}\alpha_m x^1 + \mathrm{i}\gamma_m^{(0)-} x^3), \qquad x^3 \le 0. \tag{13.52b}$$

From (13.42), (13.51), and (13.52) we obtain the **W** matrices in the same spirit as in Subsection 13.2.2.3,

$$\mathbf{W}^{(p)} = \begin{pmatrix} \delta_{mn} & \delta_{mn} \\ \dfrac{\gamma_m^{(p)+} - \alpha_m \sin\zeta}{k_0^* \varepsilon^{(p)}} \delta_{mn} & \dfrac{\gamma_m^{(p)-} - \alpha_m \sin\zeta}{k_0^* \varepsilon^{(p)}} \delta_{mn} \end{pmatrix}, \qquad p = 0, 2. \tag{13.53}$$

$$\mathbf{W}^{(1)} = \begin{pmatrix} H_{2mq}^{+} & H_{2mq}^{-} \\ E_{1mq}^{+} & E_{1mq}^{-} \end{pmatrix}. \tag{13.54}$$

These **W** matrices do not have the symmetry noted after (13.29). As illustrated in Subsection 13.2.2, after the **W** matrices are obtained, the grating problem is mostly solved. The matrix ϕ and vectors $\mathscr{F}$ and F can be defined similarly, using $\gamma_m^{(p)\pm}$ to replace $\pm\gamma_m^{(p)}$ and (x^1, x^3) to replace (x, z). After the resulting boundary matching equations are solved by using the S matrix algorithm, the Rayleigh amplitudes $R_m^{(\mathrm{h})}$ and $T_m^{(\mathrm{h})}$ are obtained. The diffraction efficiencies are given by

$$\eta_{m,\mathrm{TM}}^{(2)} = \frac{\gamma_m^{(2)+} - \alpha_m \sin\zeta}{\alpha_0 \sin\zeta - \gamma_0^{(2)-}} |R_m^{(\mathrm{h})}|^2, \qquad m \in U^{(2)}, \tag{13.55a}$$

$$\eta_{m,\mathrm{TM}}^{(0)} = \frac{\varepsilon^{(2)}}{\varepsilon^{(0)}} \cdot \frac{\alpha_m \sin\zeta - \gamma_m^{(0)-}}{\alpha_0 \sin\zeta - \gamma_0^{(2)-}} |T_m^{(\mathrm{h})}|^2, \qquad m \in U^{(0)}. \tag{13.55b}$$

13.3 One-dimensional gratings in conical mounting

13.3.1 Description of state of polarization of the incident and diffracted waves

When the wave vector $\mathbf{k}_{\mathrm{inc}}$ of an incident plane wave is in the principal plane of the grating all diffraction orders are also in the principal plane, and if the incident light is TE(TM) polarized, so are all diffracted orders. When $\mathbf{k}_{\mathrm{inc}}$ is not in the principal plane all diffracted orders fall on the side of a cone, whose axis is along the groove direction and whose cone angle is determined by the projection of $\mathbf{k}_{\mathrm{inc}}$ on the groove direction, hence the name conical mounting. It takes two angles to define $\mathbf{k}_{\mathrm{inc}}$. They are typically chosen as the polar angle θ formed by $\mathbf{k}_{\mathrm{inc}}$ with the grating normal and the azimuth angle ϕ that is the angle between the groove periodic direction and the projection of $\mathbf{k}_{\mathrm{inc}}$ on the grating plane (see Fig. 13.5 where the angle ϕ shown is positive). The ranges of θ and ϕ are $0 \le \theta < \pi/2$ and $-\pi < \phi \le \pi$, respectively. It is easy to verify that

$$\mathbf{k}_{\mathrm{inc}} = \mathbf{b}^1\alpha_0 + \mathbf{b}^2\beta_0 + \mathbf{b}^3\gamma_0^{(2)-}, \tag{13.56}$$

where

$$\alpha_0 = k^{(2)}\sin\theta\cos\phi, \quad \beta_0 = k^{(2)}\sin\theta\sin\phi, \quad \gamma_0^{(2)-} = \alpha_0\sin\zeta - k^{(2)}\cos\theta\cos\zeta. \tag{13.57}$$

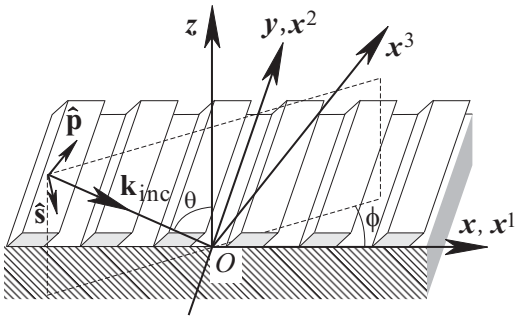


Fig. 13.5. Definition and notation of a slanted-grating problem in conical mounting.

The polarizations of the diffracted orders are in general elliptical even when the incident plane wave is linearly polarized. The polarization states of both incident and diffracted light in the two semi-infinite media can be fully specified in a number of ways. From a theoretical point of view the simplest is to use complex amplitudes of the y components of the electric and magnetic fields (the theoretical specification). From a practical point of view it is better to

define polarization with respect to a local reference frame $(\hat{\mathbf{p}}, \hat{\mathbf{s}}, \mathbf{k})$ that forms an orthogonal triplet with

$$\hat{\mathbf{s}} = \frac{\mathbf{k} \times \hat{\mathbf{z}}}{|\mathbf{k} \times \hat{\mathbf{z}}|}, \qquad \hat{\mathbf{p}} = \frac{\hat{\mathbf{s}} \times \mathbf{k}}{|\hat{\mathbf{s}} \times \mathbf{k}|}, \tag{13.58}$$

where $\mathbf{k}$ stands for the wave vector of the incident or a diffracted, reflected or transmitted plane wave. Due to transversality of the electromagnetic field of a plane wave in a homogeneous isotropic medium, the electric field vector of the incident or a diffracted order has this decomposition:

$$\mathbf{E} = (\mathbf{E} \cdot \hat{\mathbf{s}})\hat{\mathbf{s}} + (\mathbf{E} \cdot \hat{\mathbf{p}})\hat{\mathbf{p}} = E_{\mathrm{s}}\hat{\mathbf{s}} + E_{\mathrm{p}}\hat{\mathbf{p}}. \tag{13.59}$$

So, it suffices to know E_{s} and E_{p}, which are in general complex (the analytical specification). A more geometrical method is to use two angles (the analytic-geometric specification),

$$\alpha = \arctan\frac{|E_{\mathrm{s}}|}{|E_{\mathrm{p}}|}, \qquad \delta = \arg\frac{E_{\mathrm{p}}}{E_{\mathrm{s}}}. \tag{13.60}$$

The fully geometric specification is to use the orientation angle ψ of the major axis of the polarization ellipse with respect to the p axis and the ellipticity angle

$$\chi = \pm \arctan\frac{b}{a}, \tag{13.61}$$

where the a and b are the lengths the major and minor axes of the polarization ellipse, respectively, and the ± signs distinguish between right-handed and left-handed waves. It is elementary to convert from one specification to another one, typically from analytic-geometric or geometric one to theoretic one for the incident plane wave and in the reverse direction for the calculated diffracted waves.

In any case, it is recommended to use the (p, s) reference to describe polarization state for conical mountings and to reserve TE or TM for non-conical mounting.

13.3.2 Isotropic gratings

The Fourier modal method has been applied by several authors to treat rectangular gratings in conical mounting, but to date a complete and fully correct formulation is still unpublished. The method has also been applied to slanted volume gratings or parallelogram gratings [13.15, 13.16]; however, the presented formulations are inefficient because the rectangular Cartesian coordinate system is used and the size of the derived matrix eigenvalue problem is twice as large as it could be. In this section we continue to deal with slanted gratings. An efficient formulation that uses the minimum matrix size is derived in the matched oblique coordinate system. A full set of formulas for the un-slanted grating problem are given as a special case.

13.3.2.1 Rayleigh expansions in oblique coordinates and conical mounting

In conical mounting Rayleigh expansions are still available in the two semi-infinite homogenous media. Since the grating structure is invariant in the y direction, an exponential y dependence is shared by all field components and it is convenient to work with the y components of the fields. It is easy to show that

$$E_2(x^1, x^2, x^3) = I^{(\mathrm{e})} \exp[\mathrm{i}(\alpha_0 x^1 + \beta_0 x^2 + \gamma_0^{(2)-} x^3)] + \\ + \sum_{m=-\infty}^{+\infty} R_m^{(\mathrm{e})} \exp[\mathrm{i}(\alpha_m x^1 + \beta_0 x^2 + \gamma_m^{(2)+} x^3)], \qquad x^3 \geq \tilde{h}; \tag{13.62a}$$

$$E_2(x^1,x^2,x^3)=\sum_{m=-\infty}^{+\infty} T_m^{(\mathrm{e})}\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(0)-}x^3)], \qquad x^3\le 0. \tag{13.62b}$$

$$H_2(x^1,x^2,x^3)=I^{(\mathrm{h})}\exp[\mathrm{i}(\alpha_0 x^1+\beta_0 x^2+\gamma_0^{(2)-}x^3)]+ \\ +\sum_{m=-\infty}^{+\infty} R_m^{(\mathrm{h})}\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(2)+}x^3)], \qquad x^3\ge \tilde{h}; \tag{13.63a}$$

$$H_2(x^1,x^2,x^3)=\sum_{m=-\infty}^{+\infty} T_m^{(\mathrm{h})}\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(0)-}x^3)], \qquad x^3\le 0. \tag{13.63b}$$

where

$$\gamma_m^{(p)\pm}=\alpha_m\sin\zeta\pm\sqrt{\tilde{k}^{(p)2}-\alpha_m^2}\,\cos\zeta, \\ \tilde{k}^{(p)2}=k^{(p)2}-\beta_0^2, \qquad \mathrm{Re}[\sqrt{\tilde{k}^{(p)2}-\alpha_m^2}\,]+\mathrm{Im}[\sqrt{\tilde{k}^{(p)2}-\alpha_m^2}\,]>0, \qquad p=0,2, \tag{13.64}$$

and $I^{(\mathrm{e})}$ and $I^{(\mathrm{h})}$ are the y components of the incident electric and magnetic vectors, respectively. Evidently here $\tilde{k}^{(p)2}$ plays the role that $k^{(p)2}$ plays in (13.43). The arguments of exponential functions in the Rayleigh expansions can be written as

$$\mathrm{i}\mathbf{k}_m^{(p)\pm}\cdot\mathbf{r}=\mathrm{i}(\alpha_m\mathbf{b}^1+\beta_0\mathbf{b}^2+\gamma_m^{(p)\pm}\mathbf{b}^3)\cdot(\mathbf{b}_1x^1+\mathbf{b}_2x^2+\mathbf{b}_3x^3). \tag{13.65}$$

Therefore, α_m, β_0, and $\gamma_m^{(p)\pm}$ are the first, second, and third covariant components of the mth diffracted wave vector $\mathbf{k}_m^{(p)\pm}$, respectively.

13.3.2.2 Minimum-matrix-size eigenvalue problem

The key to achieving a minimum-matrix-size formulation is to make decomposition of the fields in the periodic layer according to two characteristic polarization states. Since the modes of the periodic region are sought as if the region is infinite in the vertical direction, the geometric modal picture of Fig. 13.2 is still valid. As is remarked a few lines below (13.50) a rotation of the picture about the y axis to fit the coordinate system of Fig. 13.5 is needed. However, a major difference takes place here. In conical mounting the plane of the zigzag paths of a mode is still perpendicular to the slanted medium interfaces but it is no longer perpendicular to the y axis. It is rotated around the vector $\mathbf{b}^1$ of Fig. 13.4 by an angle such that the projection of the wave vectors on the y axis is β_0. Since β_0 is a constant and the zigzag angles depend on both the mode order number and polarization, so does this rotation angle. Therefore, a nonzero β_0 results in two sets of infinitely many planes of zigzag paths. Within one set the electric field vector of a mode is along the normal of its associated zigzag plane, and within the other set the magnetic field vector is along the normal. Although the normal vectors of the zigzag planes are different, they have one feature in common: they are all perpendicular to $\mathbf{b}^1$. If we recall the geometric interpretation of contravariant components of a vector, we can say that in the case of conical mounting the modes within the periodic region are characterized by $E^1 = 0$ and $H^1 = 0$. A mode with $E^1 = 0$ will be called an $E_\perp$ mode and denoted by a superscript (e), and a mode with $H^1 = 0$ will be called an $H_\perp$ mode and denoted by a superscript (h). In a conical diffraction problem both polarization modes are excited and any field inside the periodic medium can be decomposed into a superposition of $E_\perp$ modes and $H_\perp$ modes.

From Maxwell equations it can be shown that in a medium with ε such that $\partial_2\varepsilon = \partial_3\varepsilon = 0$, the second covariant component of magnetic field vector of an $H_\perp$ mode obeys equation

$$\varepsilon\,\partial_\rho \frac{1}{\varepsilon} g^{\rho\sigma}\partial_\sigma H_2^{(\mathrm{h})} + \varepsilon\mu k_0^2 H_2^{(\mathrm{h})} = 0. \tag{13.66}$$

After expansion and rearrangement for the purpose of Fourier factorization it becomes

$$\varepsilon\left[\mu k_0^2 + \frac{1}{\cos^2\zeta}(\partial_1 - \sin\zeta\,\partial_3)\frac{1}{\varepsilon}(\partial_1 - \sin\zeta\,\partial_3)\right]H_2^{(\mathrm{h})} + (\partial_2^2 + \partial_3^2)H_2^{(\mathrm{h})} = 0. \tag{13.67}$$

In the next subsection we will see that $(1/\varepsilon)(\partial_1 - \sin\zeta\,\partial_3)H_2$ is proportional to E_3; therefore, similar to the treatment of (13.17), for both appearances of ε in (13.67) the inverse rule should be used. With an exponential dependence $\exp(\mathrm{i}\beta_0 x^2 + \mathrm{i}\gamma x^3)$ for the modal fields, (13.67) becomes

$$\begin{aligned}&\left[\cos^2\zeta\left(\mu k_0^2\mathbf{I} - \beta_0^2\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]\right) - \boldsymbol{\alpha}[\![\varepsilon]\!]^{-1}\boldsymbol{\alpha}\right]H_2^{(\mathrm{h})} + \\ &+ \gamma\sin\zeta\left([\![\varepsilon]\!]^{-1}\boldsymbol{\alpha} + \boldsymbol{\alpha}[\![\varepsilon]\!]^{-1}\right)H_2^{(\mathrm{h})} - \gamma^2\left(\cos^2\zeta\left[\!\left[\frac{1}{\varepsilon}\right]\!\right] + \sin^2\zeta\,[\![\varepsilon]\!]^{-1}\right)H_2^{(\mathrm{h})} = 0,\end{aligned} \tag{13.68}$$

which is a quadratic equation in γ. Suppose the column vector $H_2^{(\mathrm{h})}$ is truncated to be $N \times 1$, then to turn the equation into a standard matrix eigenvalue problem it is necessary to double the length of the eigenvector. There are a number of ways to do this and perhaps the simplest way is to let $(\gamma H_2^{(\mathrm{h})}, H_2^{(\mathrm{h})})^{\mathrm{T}}$ be the enlarged vector. Then (13.68) is equivalent to

$$\begin{aligned}&\begin{pmatrix}\mathbf{Q}\sin\zeta\left(\boldsymbol{\alpha}[\![\varepsilon]\!]^{-1} + [\![\varepsilon]\!]^{-1}\boldsymbol{\alpha}\right) & \mathbf{Q}\left[\cos^2\zeta\left(\mu k_0^2\mathbf{I} - \beta_0^2\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]\right) - \boldsymbol{\alpha}[\![\varepsilon]\!]^{-1}\boldsymbol{\alpha}\right]\\ \mathbf{I} & 0\end{pmatrix}\begin{pmatrix}\gamma H_2^{(\mathrm{h})}\\ H_2^{(\mathrm{h})}\end{pmatrix}\\ &= \gamma\begin{pmatrix}\gamma H_2^{(\mathrm{h})}\\ H_2^{(\mathrm{h})}\end{pmatrix}, \qquad \mathbf{Q} = \left(\cos^2\zeta\left[\!\left[\frac{1}{\varepsilon}\right]\!\right] + \sin^2\zeta\,[\![\varepsilon]\!]^{-1}\right)^{-1}.\end{aligned} \tag{13.69}$$

The matrix eigenvalue problem for an $E_\perp$ mode can be obtained by using the electromagnetic symmetry. Making the change $H_2^{(\mathrm{h})} \to E_2^{(\mathrm{e})}$ and exchange $\varepsilon \leftrightarrow -\mu$ in (13.69) and using the fact that μ is a constant we get

$$\begin{pmatrix}2\sin\zeta\,\boldsymbol{\alpha} & \cos^2\zeta\,(\mu k_0^2[\![\varepsilon]\!] - \beta_0^2) - \boldsymbol{\alpha}^2\\ \mathbf{I} & 0\end{pmatrix}\begin{pmatrix}\gamma E_2^{(\mathrm{e})}\\ E_2^{(\mathrm{e})}\end{pmatrix} = \gamma\begin{pmatrix}\gamma E_2^{(\mathrm{e})}\\ E_2^{(\mathrm{e})}\end{pmatrix}. \tag{13.70}$$

The use of $[\![\varepsilon]\!]$ in the above is justified because ε is multiplied by $E_2^{(\mathrm{e})}$, a continuous quantity. The size of the matrix eigenvalue problems in (13.69) and (13.70) are both $2N \times 2N$. The computation time to solve each of them is about 1/8 of the time needed to solve the $4N \times 4N$ matrix eigenvalue problems in the previously published formulations of the slanted gratings in conical mountings. Since both $E_\perp$ and $H_\perp$ modes are needed for matching boundary conditions, the overall time saving factor is 1/4.

The eigenvalue spectra $\sigma^{(\mathrm{e})}$ of the $E_\perp$ modes and $\sigma^{(\mathrm{h})}$ of the $H_\perp$ modes are to be partitioned separately in the same way as explained in Subsection 13.2.3.3. Besides the superscripts ± and subscript $q = 1, 2, \ldots, N$, the eigenvalues will be superscripted with (e) and (h). Altogether we have $4N$ eigen-solutions. Note that for a fixed q, $\gamma_q^{(\mathrm{e})\pm}$ is not related to $\gamma_q^{(\mathrm{h})\pm}$ although they have the same subscript q.

13.3.2.3 Construction of the total fields

In a conical diffraction problem four vector components of the electromagnetic fields, E_2, H_2, E_1, and H_1, are needed to match the boundary conditions between two adjacent spatial regions. Both $E_\perp$ and $H_\perp$ modes contribute to these four field components, so we need eight sub-components: $E_2^{(\mathrm{e})}$, $H_2^{(\mathrm{h})}, E_2^{(\mathrm{h})}$, $H_2^{(\mathrm{e})}, E_1^{(\mathrm{e})}$, $H_1^{(\mathrm{h})}, E_1^{(\mathrm{h})}$, and $H_1^{(\mathrm{e})}$. The first two are already given as solutions to (13.69) and (13.70). The others can be expressed in terms of the first two as follows.

In a waveguide or diffraction structure that is invariant along a Cartesian coordinate axis, such as the x^2 axis in Fig. 13.5, the transverse components of the electromagnetic field vectors can be expressed in terms of their longitudinal components. In our present oblique coordinate system

$$H^1 = \frac{\mathrm{i}}{\tilde{k}^2}\left(\beta_0 g^{1\tau}\partial_\tau H_2 + \frac{k_0\varepsilon}{\sqrt{g}}\partial_3 E_2\right). \tag{13.71}$$

$$E_3 = \frac{\mathrm{i}}{\tilde{k}^2}\left(\beta_0\partial_3 E_2 + k_0\mu\sqrt{g}\, g^{3\tau}\partial_\tau H_2\right). \tag{13.72}$$

Then, for any $H_\perp$ field it immediately follows that

$$\partial_3 E_2^{(\mathrm{h})} = -\frac{\beta_0}{k_0^*\varepsilon}(\partial_1 - \sin\zeta\,\partial_3) H_2^{(\mathrm{h})}. \tag{13.73}$$

Substituting (13.73) into (13.72) gives

$$E_3^{(\mathrm{h})} = \frac{\mathrm{i}}{k_0^*\varepsilon}(\partial_1 - \sin\zeta\,\partial_3) H_2^{(\mathrm{h})}. \tag{13.74}$$

Since $E_3^{(\mathrm{h})}$ is continuous Fourier transforming (13.73) leads to

$$E_{2q}^{(\mathrm{h})} = -\frac{\beta_0}{k_0^*\gamma_q^{(\mathrm{h})}}[\![\varepsilon]\!]^{-1}(\boldsymbol{\alpha} - \gamma_q^{(\mathrm{h})}\sin\zeta)\, H_{2q}^{(\mathrm{h})}. \tag{13.75a}$$

This equation is valid for a mode, as indicated by the attached subscript q to the eigenvalue and the field components. Next, $E_1^{(\mathrm{e})}$ can be obtained by using (13.47) and equation $\nabla\cdot\mathbf{D} = 0$. The former gives $E_1^{(\mathrm{e})} = \sin\zeta E_3^{(\mathrm{e})}$ and $E_3^{(\mathrm{e})} = E^{3(\mathrm{e})}$, and the latter can be expanded as

$$\varepsilon\nabla\cdot\mathbf{E}^{(\mathrm{e})} + (\nabla\varepsilon)\cdot\mathbf{E}^{(\mathrm{e})} = \varepsilon\partial_\sigma E^{\sigma(\mathrm{e})} + (\partial_1\varepsilon)E^{1(\mathrm{e})} = \partial_2 E^{2(\mathrm{e})} + \partial_3 E^{3(\mathrm{e})} = \partial_2 E_2^{(\mathrm{e})} + \partial_3 E_3^{(\mathrm{e})} = 0. \tag{13.76}$$

It then follows that

$$E_{1q}^{(\mathrm{e})} = -\frac{\beta_0}{\gamma_q^{(\mathrm{e})}}\sin\zeta\, E_{2q}^{(\mathrm{e})}. \tag{13.77a}$$

Finally, $E_1^{(\mathrm{h})}$ can be obtained by using (13.45a) and (13.74). The result is

$$E_{1q}^{(\mathrm{h})} = \frac{1}{k_0^*\gamma_q^{(\mathrm{h})}}\left[\mu k_0^2\cos^2\zeta - \boldsymbol{\alpha}[\![\varepsilon]\!]^{-1}(\boldsymbol{\alpha} - \gamma_q^{(\mathrm{h})}\sin\zeta)\right] H_{2q}^{(\mathrm{h})}. \tag{13.78a}$$

From the electromagnetic symmetry the magnetic images of (13.75a), (13.77a) and (13.78a) are

$$H_{2q}^{(\mathrm{e})} = \frac{\beta_0}{k_0^* \mu \gamma_q^{(\mathrm{e})}} (\boldsymbol{\alpha} - \gamma_q^{(\mathrm{e})} \sin\zeta)\, E_{2q}^{(\mathrm{e})}, \tag{13.75b}$$

$$H_{1q}^{(\mathrm{h})} = -\frac{\beta_0}{\gamma_q^{(\mathrm{h})}} \sin\zeta\, H_{2q}^{(\mathrm{h})}, \tag{13.77b}$$

$$H_{1q}^{(\mathrm{e})} = \frac{-1}{k_0^* \mu \gamma_q^{(\mathrm{e})}} \Big[\mu k_0^2 \cos^2\zeta \, [\![\varepsilon]\!] - \boldsymbol{\alpha} (\boldsymbol{\alpha} - \gamma_q^{(\mathrm{e})} \sin\zeta) \Big] E_{2q}^{(\mathrm{e})}. \tag{13.78b}$$

Collecting all of the above expressions of the sub-components of the fields we have

$$\mathscr{F}(x^3) = \mathbf{W}^{(1)}\, \phi(x^3)\, F, \tag{13.79}$$

where

$$\mathscr{F} = (E_{2m}, H_{2m}, H_{1m}, E_{1m})^{\mathrm{T}}, \qquad F = (\tilde{u}_q^{(\mathrm{e})}, \tilde{u}_q^{(\mathrm{h})}, \tilde{d}_q^{(\mathrm{e})}, \tilde{d}_q^{(\mathrm{h})})^{\mathrm{T}}, \tag{13.80}$$

$$\mathbf{W}^{(1)} = \begin{pmatrix} E_{2mq}^{(\mathrm{e})+} & -\beta_0 V_{mq}^{(\mathrm{h})+} & E_{2mq}^{(\mathrm{e})-} & -\beta_0 V_{mq}^{(\mathrm{h})-} \\ \beta_0 V_{mq}^{(\mathrm{e})+} & H_{2mq}^{(\mathrm{h})+} & \beta_0 V_{mq}^{(\mathrm{e})-} & H_{2mq}^{(\mathrm{h})-} \\ -U_{mq}^{(\mathrm{e})+} & -B_q^{(\mathrm{h})+} H_{2mq}^{(\mathrm{h})+} & -U_{mlq}^{(\mathrm{e})-} & -B_q^{(\mathrm{h})-} H_{2mq}^{(\mathrm{h})-} \\ -B_q^{(\mathrm{e})+} E_{2mq}^{(\mathrm{e})+} & U_{mq}^{(\mathrm{h})+} & -B_q^{(\mathrm{e})-} E_{2mq}^{(\mathrm{e})-} & U_{mlq}^{(\mathrm{h})-} \end{pmatrix}, \tag{13.81}$$

$$\phi^{(1)} = \begin{pmatrix} \exp[\mathrm{i}\gamma_q^{(\mathrm{e})+}(x^3 - x_0^3)] & 0 & 0 & 0 \\ 0 & \exp[\mathrm{i}\gamma_q^{(\mathrm{h})+}(x^3 - x_0^3)] & 0 & 0 \\ 0 & 0 & \exp[\mathrm{i}\gamma_q^{(\mathrm{e})-}(x^3 - x_1^3)] & 0 \\ 0 & 0 & 0 & \exp[\mathrm{i}\gamma_q^{(\mathrm{h})-}(x^3 - x_1^3)] \end{pmatrix}. \tag{13.82}$$

Note that in (13.80) H_{1m} is listed before E_{1m} in $\mathscr{F}$ and the modal amplitudes in F are phase-adjusted in a way similar to that in (13.24) and (13.32). In (13.81)

$$V_{mq}^{(\mathrm{e})\pm} = \frac{1}{k_0^* \mu \gamma_q^{(\mathrm{e})\pm}} \left(\alpha_m - \gamma_q^{(\mathrm{e})\pm} \sin\zeta \right) E_{2mq}^{(\mathrm{e})\pm}, \tag{13.83a}$$

$$V_{mq}^{(\mathrm{h})\pm} = \frac{1}{k_0^* \gamma_q^{(\mathrm{h})\pm}} \sum_{n=m_1}^{m_2} [\![\varepsilon]\!]_{mn}^{-1} \left(\alpha_n - \gamma_q^{(\mathrm{h})\pm} \sin\zeta \right) H_{2nq}^{(\mathrm{h})\pm}, \tag{13.83b}$$

$$U_{mq}^{(\mathrm{e})\pm} = \frac{k_0^*}{\gamma_q^{(\mathrm{e})\pm}} \sum_{n=m_1}^{m_2} [\![\varepsilon]\!]_{mn} E_{2nq}^{(\mathrm{e})\pm} - \alpha_m V_{mq}^{(\mathrm{e})\pm}, \tag{13.84a}$$

$$U_{mq}^{(\mathrm{h})\pm} = \frac{k_0^* \mu}{\gamma_q^{(\mathrm{h})\pm}} H_{2mq}^{(\mathrm{h})\pm} - \alpha_m V_{mq}^{(\mathrm{h})\pm}, \tag{13.84b}$$

$B_q^{(s)\pm} = \beta_0 \sin\zeta / \gamma_q^{(s)\pm}$, and s = e, h. Using (13.69) and (13.70), (13.84) can be written in another way

$$U_{mq}^{(\mathrm{e})\pm} = \frac{1}{k_0^* \mu \gamma_q^{(\mathrm{e})\pm}} \Big[\cos^2\zeta\, \beta_0^2 + (\gamma_q^{(\mathrm{e})\pm})^2 - \gamma_q^{(\mathrm{e})\pm} \sin\zeta\, \alpha_m \Big] E_{2mq}^{(\mathrm{e})\pm}, \tag{13.85a}$$

$$U_{mq}^{(\mathrm{h})\pm}=\frac{1}{k_0^*\gamma_q^{(\mathrm{h})\pm}}\sum_{n=m_1}^{m_2}\left[\left(\cos^2\zeta\,\beta_0^2\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]+(\gamma_q^{(\mathrm{h})\pm})^2\mathbf{Q}^{-1}\right)_{mn}-\gamma_q^{(\mathrm{h})\pm}\sin\zeta\left(\left[\!\left[\varepsilon\right]\!\right]^{-1}\right)_{mn}\alpha_n\right]H_{2nq}^{(\mathrm{h})\pm}. \quad (13.85b)$$

A comparison between (13.84) and (13.85) shows that it is simpler to use (13.85a) to compute $U_{mq}^{(\mathrm{e})\pm}$ and to use (13.84b) to compute $U_{mq}^{(\mathrm{h})\pm}$.

From (13.62) and (13.63) the Rayleigh expansions of E_1 and H_1 in the homogeneous regions can be derived from Maxwell equations,

$$E_1=[-\tau_3^{(2)}\alpha_0 I^{(\mathrm{e})}+\tau_2^{(2)}(\gamma_0^{(2)-}-\alpha_0\sin\zeta)I^{(\mathrm{h})}]\exp[\mathrm{i}(\alpha_0 x^1+\beta_0 x^2+\gamma_0^{(2)-}x^3)]+$$
$$+\sum_{m=-\infty}^{+\infty}[-\tau_3^{(2)}\alpha_m R_m^{(\mathrm{e})}+\tau_2^{(2)}(\gamma_m^{(2)+}-\alpha_m\sin\zeta)R_m^{(\mathrm{h})}]\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(2)+}x^3)],$$
$$x^3\ge\tilde{h}; \quad (13.86a)$$

$$E_1=\sum_{m=-\infty}^{+\infty}[-\tau_3^{(0)}\alpha_m T_m^{(\mathrm{e})}+\tau_2^{(0)}(\gamma_m^{(0)-}-\alpha_m\sin\zeta)T_m^{(\mathrm{h})}]\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(0)-}x^3)],$$
$$x^3\le 0; \quad (13.86b)$$

$$H_1=[-\tau_3^{(2)}\alpha_0 I^{(\mathrm{h})}-\tau_1^{(2)}(\gamma_0^{(2)-}-\alpha_0\sin\zeta)I^{(\mathrm{e})}]\exp[\mathrm{i}(\alpha_0 x^1+\beta_0 x^2+\gamma_0^{(2)-}x^3)]+$$
$$+\sum_{m=-\infty}^{+\infty}[-\tau_3^{(2)}\alpha_m R_m^{(\mathrm{h})}-\tau_1^{(2)}(\gamma_m^{(2)+}-\alpha_m\sin\zeta)R_m^{(\mathrm{e})}]\exp[\mathrm{i}(\alpha_m x^1+\beta_0 x^2+\gamma_m^{(2)+}x^3)],$$
$$x^3\ge\tilde{h}; \quad (13.87a)$$

$$H_1=\sum_{m=-\infty}^{+\infty}[-\tau_3^{(0)}\alpha_m T_m^{(\mathrm{h})}-\tau_1^{(0)}(\gamma_m^{(0)-}-\alpha_m\sin\zeta)T_m^{(\mathrm{e})}]\exp(\mathrm{i}\alpha_m x^1+\mathrm{i}\beta_0 x^2+\mathrm{i}\gamma_m^{(0)-}x^3),$$
$$x^3\le 0, \quad (13.87b)$$

where

$$\tau_1^{(p)}=\frac{k_0\varepsilon^{(p)}}{\tilde{k}^{(p)2}\cos\zeta}, \qquad \tau_2^{(p)}=\frac{k_0\mu^{(p)}}{\tilde{k}^{(p)2}\cos\zeta}, \qquad \tau_3^{(p)}=\frac{\beta_0}{\tilde{k}^{(p)2}}. \quad (13.88)$$

With these results we can write down the counterparts of the right-hand side of (13.79) for the homogeneous regions:

$$F^{(2)}=(\tilde{R}_m^{(\mathrm{e})},\tilde{R}_m^{(\mathrm{h})},\tilde{I}^{(\mathrm{e})}\delta_{m0},\tilde{I}^{(\mathrm{h})}\delta_{m0})^{\mathrm{T}}, \qquad F^{(0)}=(0,0,\tilde{T}_m^{(\mathrm{e})},\tilde{T}_m^{(\mathrm{h})})^{\mathrm{T}}, \quad (13.89)$$

$$\mathbf{W}^{(p)}=\begin{pmatrix}\delta_{mn} & 0 & \delta_{mn} & 0\\ 0 & \delta_{mn} & 0 & \delta_{mn}\\ -\tau_1^{(p)}\Gamma_m^{(p)+}\delta_{mn} & -\tau_3^{(p)}\alpha_m\delta_{mn} & -\tau_1^{(p)}\Gamma_m^{(p)-}\delta_{mn} & -\tau_3^{(p)}\alpha_m\delta_{mn}\\ -\tau_3^{(p)}\alpha_m\delta_{mn} & \tau_2^{(p)}\Gamma_m^{(p)+}\delta_{mn} & -\tau_3^{(p)}\alpha_m\delta_{mn} & \tau_2^{(p)}\Gamma_m^{(p)-}\delta_{mn}\end{pmatrix}, \quad p=0,2, \quad (13.90)$$

$$\phi^{(p)}=\begin{pmatrix}\exp(\mathrm{i}\gamma_m^{(p)+}x^3) & 0 & 0 & 0\\ 0 & \exp(\mathrm{i}\gamma_m^{(p)+}x^3) & 0 & 0\\ 0 & 0 & \exp(\mathrm{i}\gamma_m^{(p)-}x^3) & 0\\ 0 & 0 & 0 & \exp(\mathrm{i}\gamma_m^{(p)-}x^3)\end{pmatrix}, \quad p=0,2, \quad (13.91)$$

where $\Gamma_m^{(p)\pm} = \gamma_m^{(p)\pm} - \alpha_m \sin\zeta$, and $\tilde{R}_m^{(\mathrm{e})}, \tilde{R}_m^{(\mathrm{h})}$, etc. are phase adjusted as in (13.25).

Using (13.62), (13.63), (13.86), and (13.87), from the basic definition of diffraction efficiency one can derive the diffraction efficiencies of reflected and transmitted orders in conical mounting

$$\eta_m^{(2)} = \frac{\gamma_m^{(2)+} - \alpha_m \sin\zeta}{\alpha_0 \sin\zeta - \gamma_0^{(2)-}} \cdot \frac{\varepsilon^{(2)} |R_m^{(\mathrm{e})}|^2 + \mu |R_m^{(\mathrm{h})}|^2}{\varepsilon^{(2)} |I^{(\mathrm{e})}|^2 + \mu |I^{(\mathrm{h})}|^2}, \qquad m \in U^{(2)}, \tag{13.92a}$$

$$\eta_m^{(0)} = \frac{\tilde{k}^{(2)2}}{\tilde{k}^{(0)2}} \cdot \frac{\alpha_m \sin\zeta - \gamma_m^{(0)-}}{\alpha_0 \sin\zeta - \gamma_0^{(2)-}} \cdot \frac{\varepsilon^{(0)} |T_m^{(\mathrm{e})}|^2 + \mu |T_m^{(\mathrm{h})}|^2}{\varepsilon^{(2)} |I^{(\mathrm{e})}|^2 + \mu |I^{(\mathrm{h})}|^2}, \qquad m \in U^{(0)}. \tag{13.92b}$$

The two Rayleigh amplitudes $R_m^{(\mathrm{e})}$ and $R_m^{(\mathrm{h})}$ fully specify the state of polarization of the mth reflected order. To convert this specification to the analytic-geometric specification one can take the following steps. Since $R_m^{(\mathrm{e})}$ stands for E_2 and E_1 is already given in (13.86), to complete the information about **E**, E_3 can be obtained by using transversality of plane wave, $\mathbf{k} \cdot \mathbf{E} = k_\rho g^{\rho\sigma} E_\sigma = 0$, where $k_1 = \alpha_m$, $k_2 = \beta_0$, and $k_3 = \gamma_m^{(p)\pm}$. The local reference frame $(\hat{\mathbf{p}}, \hat{\mathbf{s}}, \mathbf{k})$ can be established once **k** is known. Then, E_s and E_p can be easily obtained from (13.58) and (13.59).

The previous formulations of slanted gratings in conical mounting output all four tangential components of the electromagnetic fields needed for matching boundary conditions as eigenvectors of a $4N \times 4N$ eigenvalue problem, so the **W** matrix is built automatically except sorting of eigenvalues. In the present formulation after the eigenvectors are obtained some work is needed to build the **W** matrix. However, this additional work is insignificant compared with the saving gained by solving two $2N \times 2N$ eigenvalue problems.

13.3.2.4 Special cases of rectangular gratings and non-conical mounting

For rectangular isotropic gratings in conical mounting all required formulas can be obtained by setting $\zeta = 0$ in Subsections 13.3.2.1-3.2.3. In particular, it follows from (13.69) and (13.70) the $H_\perp$ and $E_\perp$ modes can be solved from two simpler equations

$$\left[\!\left[\frac{1}{\varepsilon}\right]\!\right]^{-1} \left(k_0^2 \mu \mathbf{I} - \boldsymbol{\alpha} [\![\varepsilon]\!]^{-1} \boldsymbol{\alpha} \right) H_2^{(\mathrm{h})} = (\gamma^2 + \beta_0^2) H_2^{(\mathrm{h})}, \tag{13.93a}$$

$$\left(k_0^2 \mu [\![\varepsilon]\!] - \boldsymbol{\alpha}^2 \right) E_2^{(\mathrm{e})} = (\gamma^2 + \beta_0^2) E_2^{(\mathrm{e})}. \tag{13.93b}$$

The difference between (13.93a) and (13.18) is only in the eigenvalues. The $\mathbf{W}^{(1)}$ matrix for the grating layer becomes

$$\mathbf{W}^{(1)} = \begin{pmatrix} E_{2mq}^{(\mathrm{e})+} & -\tilde{V}_{mq}^{(\mathrm{h})+} & E_{2mq}^{(\mathrm{e})-} & -\tilde{V}_{mq}^{(\mathrm{h})-} \\ \tilde{V}_{mq}^{(\mathrm{e})+} & H_{2mq}^{(\mathrm{h})+} & \tilde{V}_{mq}^{(\mathrm{e})-} & H_{2mq}^{(\mathrm{h})-} \\ -\tilde{U}_{mq}^{(\mathrm{e})+} & 0 & -\tilde{U}_{mq}^{(\mathrm{e})-} & 0 \\ 0 & \tilde{U}_{mq}^{(\mathrm{h})+} & 0 & \tilde{U}_{mlq}^{(\mathrm{h})-} \end{pmatrix}, \tag{13.94}$$

where

$$\tilde{V}_{mq}^{(\mathrm{e})\pm} = \frac{\beta_0 \alpha_m}{k_0 \mu \gamma_q^{(\mathrm{e})\pm}} E_{2mq}^{(\mathrm{e})\pm}, \qquad \tilde{U}_{mq}^{(\mathrm{e})\pm} = \frac{\beta_0^2 + (\gamma_q^{(\mathrm{e})\pm})^2}{k_0 \mu \gamma_q^{(\mathrm{e})\pm}} E_{2mq}^{(\mathrm{e})\pm}, \tag{13.95a}$$

$$\tilde{V}_{mq}^{(\mathrm{h})\pm} = \frac{\beta_0}{k_0 \gamma_q^{(\mathrm{h})\pm}} \sum_{n=m_1}^{m_2} [\![\varepsilon]\!]_{mn}^{-1} \alpha_n H_{2nq}^{(\mathrm{h})\pm}, \quad \tilde{U}_{mq}^{(\mathrm{h})\pm} = \frac{\beta_0^2 + (\gamma_q^{(\mathrm{h})\pm})^2}{k_0 \gamma_q^{(\mathrm{h})\pm}} \sum_{n=m_1}^{m_2} \left[\!\!\left[\frac{1}{\varepsilon}\right]\!\!\right]_{mn} H_{2nq}^{(\mathrm{h})\pm}. \tag{13.95b}$$

The $\mathbf{W}^{(p)}$ matrices for $p = 0$ and 2 can be obtained from (13.90) by simply replacing $\Gamma_m^{(p)\pm}$ with $\gamma_m^{(p)\pm}$. Equations (13.94) and (13.95) can be compared with equations (61-65) of [13.3]. Evidently, to prepare Fourier coefficient of the total fields needed for matching boundary conditions between the periodic region and the homogeneous regions, the present formulation requires much less matrix computation.

By setting $\beta_0 = 0$ in (13.69), (13.70), (13.81), and (13.90) we obtain an alternative formulation of the Fourier modal method for slanted gratings in classical mounting. The eigenvalue problems for both TM and TE polarizations are still $2N \times 2N$, as long as the slant angle ζ is nonzero. After row exchange $2 \leftrightarrow 3$ and column exchange $2 \leftrightarrow 3$, the 4×4 $\mathbf{W}^{(p)}$ matrices become 2×2 block-diagonal; therefore, TM and TE polarizations are uncoupled. Of course this formulation is equivalent to that developed in Subsection 13.2.3. The latter formulation is apparently more efficient because E_1 is output directly by the eigenvalue-problem solver and meanwhile the amounts of matrix computation to construct the coefficient matrices of the eigenvalue problems in both formulations are about the same.

13.3.3 Anisotropic gratings

In this subsection we continue to assume the grating profile is slanted and the diffraction mounting is conical. Both the incident region and the emergent region are isotropic, but the periodic region is anisotropic. This is the most commonly encountered situation of anisotropic gratings. However, to achieve maximum generality without sacrificing clarity (in fact, to enhance clarity) we assume that the medium anisotropy is in both permittivity and permeability. Allowing anisotropic permeability preserves the formal electromagnetic symmetry of the derived formulas and helps their derivation. Figure 13.5 also depicts the present diffraction problem. The Rayleigh expansions given by (13.62), (13.63), (13.86), and (13.87) are still valid. We only need to find expressions of the total fields in the periodic layer.

For additional readings on the Fourier modal method for one-dimensional periodic gratings made with anisotropic materials the reader may consult references [13.17-21].

13.3.3.1 Fourier factorization of constitutive relations

In an anisotropic medium the two Maxwell equations containing curl operators written in the oblique Cartesian coordinate system of (13.35a) are

$$\xi^{\rho\sigma\tau} \partial_\sigma E_\tau = \mathrm{i} k_0 \sqrt{g}\, \mu^{\rho\sigma} H_\sigma, \tag{13.96a}$$

$$\xi^{\rho\sigma\tau} \partial_\sigma H_\tau = -\mathrm{i} k_0 \sqrt{g}\, \varepsilon^{\rho\sigma} E_\sigma. \tag{13.96b}$$

The permittivity tensor $\varepsilon^{\rho\sigma}$ and permeability tensor $\mu^{\rho\sigma}$ are related to their counterparts $\overline{\varepsilon}^{\rho\sigma}$ and $\overline{\mu}^{\rho\sigma}$ in the normal rectangular Cartesian system by tensor transformation

$$\varepsilon^{\rho\sigma} = \frac{\partial x^\rho}{\partial \overline{x}^{\rho'}} \frac{\partial x^\sigma}{\partial \overline{x}^{\sigma'}} \overline{\varepsilon}^{\rho'\sigma'}, \qquad \mu^{\rho\sigma} = \frac{\partial x^\rho}{\partial \overline{x}^{\rho'}} \frac{\partial x^\sigma}{\partial \overline{x}^{\sigma'}} \overline{\mu}^{\rho'\sigma'}, \tag{13.97}$$

where $\overline{x}^1 = x$, $\overline{x}^2 = y$, and $\overline{x}^3 = z$. Note that the tensor characteristics of $\varepsilon^{\rho\sigma}$ and $\mu^{\rho\sigma}$ come from two sources: the tensor characteristics of $\overline{\varepsilon}^{\rho\sigma}$ and $\overline{\mu}^{\rho\sigma}$ and the coordinate transformation. So, the presence of either one or both results in the same mathematical complexity, although the former source is physical and the latter is mathematical.

In a one-dimensional grating problem $\varepsilon^{\rho\sigma}$ and $\mu^{\rho\sigma}$ depend only on x^1 and their functional dependences on x^1 may be discontinuous. To solve the problem by the Fourier modal method, we need to write $\mathbf{D} = \boldsymbol{\varepsilon} \cdot \mathbf{E} = \varepsilon^{\rho\sigma} E_\sigma$ and $\mathbf{B} = \boldsymbol{\mu} \cdot \mathbf{H} = \mu^{\rho\sigma} H_\sigma$ in Fourier space in a way consistent with the Fourier factorization theory. The first equation in component form can be arranged as

$$D^1 = \varepsilon^{11}\left[E_1 + \left(\frac{\varepsilon^{12}}{\varepsilon^{11}}\right)E_2 + \left(\frac{\varepsilon^{13}}{\varepsilon^{11}}\right)E_3\right], \tag{13.98a}$$

$$D^2 = \left(\frac{\varepsilon^{21}}{\varepsilon^{11}}\right)D^1 + \left(\varepsilon^{22} - \frac{\varepsilon^{21}\varepsilon^{12}}{\varepsilon^{11}}\right)E_2 + \left(\varepsilon^{23} - \frac{\varepsilon^{21}\varepsilon^{13}}{\varepsilon^{11}}\right)E_3, \tag{13.98b}$$

$$D^3 = \left(\frac{\varepsilon^{31}}{\varepsilon^{11}}\right)D^1 + \left(\varepsilon^{32} - \frac{\varepsilon^{31}\varepsilon^{12}}{\varepsilon^{11}}\right)E_2 + \left(\varepsilon^{33} - \frac{\varepsilon^{31}\varepsilon^{13}}{\varepsilon^{11}}\right)E_3. \tag{13.98c}$$

In these equations all products between field components and permittivity-element dependent functions are Type 1 or Type 2 as defined in Subsection 13.2.2.2. So, the Fourier coefficients of vector **D** in terms of those of **E** are

$$D_m^\sigma = (\varepsilon^{\sigma\tau} E_\tau)_m = \sum_n (\hat{\varepsilon}^{\sigma\tau})_{mn} E_{\tau n}, \tag{13.99a}$$

where m and n are Fourier indices, and

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1} & \left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{12}}{\varepsilon^{11}}\right]\!\right] & \left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{13}}{\varepsilon^{11}}\right]\!\right] \\ \left[\!\left[\frac{\varepsilon^{21}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1} & \left[\!\left[\frac{\varepsilon^{21}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{12}}{\varepsilon^{11}}\right]\!\right] + \left[\!\left[\varepsilon^{22} - \frac{\varepsilon^{21}\varepsilon^{12}}{\varepsilon^{11}}\right]\!\right] & \left[\!\left[\frac{\varepsilon^{21}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{13}}{\varepsilon^{11}}\right]\!\right] + \left[\!\left[\varepsilon^{23} - \frac{\varepsilon^{31}\varepsilon^{13}}{\varepsilon^{11}}\right]\!\right] \\ \left[\!\left[\frac{\varepsilon^{31}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1} & \left[\!\left[\frac{\varepsilon^{31}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{12}}{\varepsilon^{11}}\right]\!\right] + \left[\!\left[\varepsilon^{32} - \frac{\varepsilon^{31}\varepsilon^{12}}{\varepsilon^{11}}\right]\!\right] & \left[\!\left[\frac{\varepsilon^{31}}{\varepsilon^{11}}\right]\!\right]\left[\!\left[\frac{1}{\varepsilon^{11}}\right]\!\right]^{-1}\left[\!\left[\frac{\varepsilon^{13}}{\varepsilon^{11}}\right]\!\right] + \left[\!\left[\varepsilon^{33} - \frac{\varepsilon^{31}\varepsilon^{13}}{\varepsilon^{11}}\right]\!\right] \end{pmatrix}. \tag{13.100}$$

This expression of permittivity tensor $\vec{\boldsymbol{\varepsilon}}$ in Fourier space is bulky and complicated, but theoretical analysis and numerical tests have shown that it is worth the trouble because it leads to superior numerical performance for grating materials with a large permittivity contrast.

The seemingly complicated expression actually follows a simple construction rule. For any 3×3 matrix **A** with elements $A^{\rho\sigma}$, where ρ, σ = 1, 2, 3 and $A^{\rho\sigma}$ can be a number or a square matrix, provided that $(A^{\tau\tau})^{-1}$ exists for a fixed τ, we define operator $l_\tau^\pm$ by $\mathbf{Z} = l_\tau^\pm(\mathbf{A})$ with

$$Z^{\rho\sigma} = \begin{cases} (A^{\tau\tau})^{-1}, & \rho = \tau, \sigma = \tau; \\ (A^{\tau\tau})^{-1} A^{\tau\sigma}, & \rho = \tau, \sigma \neq \tau; \\ A^{\rho\tau}(A^{\tau\tau})^{-1}, & \rho \neq \tau, \sigma = \tau; \\ A^{\rho\sigma} \pm A^{\rho\tau}(A^{\tau\tau})^{-1} A^{\tau\sigma}, & \rho \neq \tau, \sigma \neq \tau. \end{cases} \tag{13.101}$$

For matrix $Z^{\rho\sigma}(x^1,x^2,x^3)$ we further define F_τ to be an operator such that $\mathbf{C} = F_\tau(\mathbf{Z})$ is a block matrix whose elements $C^{\rho\sigma}$ are Toeplitz matrices generated by the Fourier coefficients of $Z^{\rho\sigma}(x^1,x^2,x^3)$ with respect to variable x^τ. So, if matrix $Z^{\rho\sigma}$ is $p \times p$, then $C^{\rho\sigma}$ is $pm \times pm$, where m is the truncation number in Fourier space. With this apparatus, (13.100) can be written succinctly as

$$\hat{\boldsymbol{\varepsilon}} = l_1^+ F_1 l_1^-(\boldsymbol{\varepsilon}). \tag{13.102a}$$

The efficiency and preciseness of this notation is evident. This will be further demonstrated when we deal with crossed gratings in Section 13.4.

For the magnetic relations we similarly have

$$B_m^{\sigma} = (\mu^{\sigma\tau} H_{\tau})_m = \sum_n (\hat{\mu}^{\sigma\tau})_{mn} H_{\tau n}\,, \tag{13.99b}$$

$$\hat{\boldsymbol{\mu}} = l_1^{+} F_1 l_1^{-}(\boldsymbol{\mu}). \tag{13.102b}$$

13.3.3.2 Construction of the total fields

For Maxwell equations as complex as (13.96), it is advantageous to perform Fourier factorization at the very beginning. In the following we will use E_1, and H_2, etc., to denote column vectors of the Fourier coefficients of $E_1(x^1, x^2, x^3)$ and $H_2(x^1, x^2, x^3)$, respectively. The effects of applying operators ∂_1 and ∂_2 on these functions are to multiply them from the left side by $\mathrm{i}\boldsymbol{\alpha}$ and $\mathrm{i}\beta_0$, where $\boldsymbol{\alpha}$ is a diagonal matrix with elements α_n. Then in Fourier space and in expanded form (13.96) becomes

$$\boldsymbol{\alpha} E_2 - \beta_0 E_1 = k_0^{*}(\hat{\mu}^{31} H_1 + \hat{\mu}^{32} H_2 + \hat{\mu}^{33} H_3), \tag{13.103a}$$

$$\beta_0 E_3 - (\partial_3/\mathrm{i})\, E_2 = k_0^{*}(\hat{\mu}^{11} H_1 + \hat{\mu}^{12} H_2 + \hat{\mu}^{13} H_3), \tag{13.103b}$$

$$(\partial_3/\mathrm{i})\, E_1 - \boldsymbol{\alpha} E_3 = k_0^{*}(\hat{\mu}^{21} H_1 + \hat{\mu}^{22} H_2 + \hat{\mu}^{23} H_3); \tag{13.103c}$$

$$\boldsymbol{\alpha} H_2 - \beta_0 H_1 = -k_0^{*}(\hat{\varepsilon}^{31} E_1 + \hat{\varepsilon}^{32} E_2 + \hat{\varepsilon}^{33} E_3), \tag{13.104a}$$

$$\beta_0 H_3 - (\partial_3/\mathrm{i})\, H_2 = -k_0^{*}(\hat{\varepsilon}^{11} E_1 + \hat{\varepsilon}^{12} E_2 + \hat{\varepsilon}^{13} E_3), \tag{13.104b}$$

$$(\partial_3/\mathrm{i})\, H_1 - \boldsymbol{\alpha} H_3 = -k_0^{*}(\hat{\varepsilon}^{21} E_1 + \hat{\varepsilon}^{22} E_2 + \hat{\varepsilon}^{23} E_3). \tag{13.104c}$$

From (13.103a) and (13.104a), the third field components E_3 and H_3 can be expressed in terms of the first and second components:

$$E_3 = -\frac{1}{k_0^{*}}(\hat{\varepsilon}^{33})^{-1}(\boldsymbol{\alpha} H_2 - \beta_0 H_1) - (\hat{\varepsilon}^{33})^{-1}\hat{\varepsilon}^{31} E_1 - (\hat{\varepsilon}^{33})^{-1}\hat{\varepsilon}^{32} E_2\,. \tag{13.105a}$$

$$H_3 = \frac{1}{k_0^{*}}(\hat{\mu}^{33})^{-1}(\boldsymbol{\alpha} E_2 - \beta_0 E_1) - (\hat{\mu}^{33})^{-1}\hat{\mu}^{31} H_1 - (\hat{\mu}^{33})^{-1}\hat{\mu}^{32} H_2\,. \tag{13.105b}$$

Substituting (13.105) into (13.103) and (13.104) gives

$$(\partial_3/\mathrm{i})\,(E_1, E_2, H_1, H_2)^{\mathrm{T}} = \mathbf{M}\,(E_1, E_2, H_1, H_2)^{\mathrm{T}}\,, \tag{13.106}$$

where the superscript T denotes block matrix transpose,

$$\mathbf{M} = \begin{pmatrix} -\tilde{\mu}^{23}\beta_0 - \boldsymbol{\alpha}\tilde{\varepsilon}^{31} & \tilde{\mu}^{23}\boldsymbol{\alpha} - \boldsymbol{\alpha}\tilde{\varepsilon}^{32} & k_0^{*}\tilde{\mu}^{21} + \frac{1}{k_0^{*}}\boldsymbol{\alpha}\tilde{\varepsilon}^{33}\beta_0 & k_0^{*}\tilde{\mu}^{22} - \frac{1}{k_0^{*}}\boldsymbol{\alpha}\tilde{\varepsilon}^{33}\boldsymbol{\alpha} \\ \tilde{\mu}^{13}\beta_0 - \beta_0\tilde{\varepsilon}^{31} & -\tilde{\mu}^{13}\boldsymbol{\alpha} - \beta_0\tilde{\varepsilon}^{32} & -k_0^{*}\tilde{\mu}^{11} + \frac{1}{k_0^{*}}\beta_0\tilde{\varepsilon}^{33}\beta_0 & -k_0^{*}\tilde{\mu}^{12} - \frac{1}{k_0^{*}}\beta_0\tilde{\varepsilon}^{33}\boldsymbol{\alpha} \\ -k_0^{*}\tilde{\varepsilon}^{21} - \frac{1}{k_0^{*}}\boldsymbol{\alpha}\tilde{\mu}^{33}\beta_0 & -k_0^{*}\tilde{\varepsilon}^{22} + \frac{1}{k_0^{*}}\boldsymbol{\alpha}\tilde{\mu}^{33}\boldsymbol{\alpha} & -\tilde{\varepsilon}^{23}\beta_0 - \boldsymbol{\alpha}\tilde{\mu}^{31} & \tilde{\varepsilon}^{23}\boldsymbol{\alpha} - \boldsymbol{\alpha}\tilde{\mu}^{32} \\ k_0^{*}\tilde{\varepsilon}^{11} - \frac{1}{k_0^{*}}\beta_0\tilde{\mu}^{33}\beta_0 & k_0^{*}\tilde{\varepsilon}^{12} + \frac{1}{k_0^{*}}\beta_0\tilde{\mu}^{33}\boldsymbol{\alpha} & \tilde{\varepsilon}^{13}\beta_0 - \beta_0\tilde{\mu}^{31} & -\tilde{\varepsilon}^{13}\boldsymbol{\alpha} - \beta_0\tilde{\mu}^{32} \end{pmatrix}, \tag{13.107}$$

and

$$\tilde{\boldsymbol{\varepsilon}} = l_3^-(\hat{\boldsymbol{\varepsilon}}) = l_3^- l_1^+ F_1 l_1^-(\boldsymbol{\varepsilon}), \quad \tilde{\boldsymbol{\mu}} = l_3^-(\hat{\boldsymbol{\mu}}) = l_3^- l_1^+ F_1 l_1^-(\boldsymbol{\mu}). \tag{13.108}$$

In writing (13.107) we have left the scalar β_0 where it was dropped down during differentiation relative to other matrices in a product, as if β_0 were a matrix. The reason for doing so will become apparent in Section 13.4. Since **M** does not depend on x^3, the modal function has this dependence $\exp(\mathrm{i}\gamma x^3)$ on x^3 and, after the operator (∂_3/i) is replaced with γ, (13.106) becomes an eigenvalue problem.

If the Fourier-Floquet series representing each field components is truncated from m_1 to $m_2 = N + m_1 - 1$, then **M** is $4N \times 4N$. Partition of the $4N$ eigensolutions follows the same principle as described in Subsection 13.2.3.3, namely, place all eigenvalues with positive and negative imaginary parts in σ^+ and σ^-, respectively, and divide the rest arbitrarily so that σ^+ and σ^- each contains $2N$ eigenvalues. Now the index q for eigensolutions runs from 1 to $2N$. Analogous to (13.51), we have

$$\begin{pmatrix} E_1 \\ E_2 \\ H_1 \\ H_2 \end{pmatrix} = \sum_{m=m_1}^{m_2} \exp(\mathrm{i}\alpha_m x^1 + \mathrm{i}\beta_0 x^2) \sum_{q=1}^{2N} \begin{pmatrix} E_{1mq}^+ & E_{1mq}^- \\ E_{2mq}^+ & E_{2mq}^- \\ H_{1mq}^+ & H_{1mq}^- \\ H_{2mq}^+ & H_{2mq}^- \end{pmatrix} \begin{pmatrix} \exp[\mathrm{i}\gamma_q^+(x^3 - x_0^3)] & 0 \\ 0 & \exp[\mathrm{i}\gamma_q^-(x^3 - x_1^3)] \end{pmatrix} \begin{pmatrix} \tilde{u}_q \\ \tilde{d}_q \end{pmatrix}. \tag{13.109}$$

The eigenvector matrix in (13.109) is the $\mathbf{W}^{(1)}$ matrix in the S-matrix algorithm formulation. Together with the $\mathbf{W}^{(0)}$ and $\mathbf{W}^{(2)}$ matrices given by (13.90) the unknown Rayleigh amplitudes in the homogeneous regions can be solved by using the S-matrix algorithm.

13.3.3.3 Special cases

Equation (13.107) defines the most general eigenvalue problem treated so far in this chapter. By separately or simultaneously setting $\beta_0 = 0$, $\zeta = 0$, and letting $\bar{\varepsilon}^{ij}$ and $\bar{\mu}^{ij}$ to be scalars or to take a special orientation, eigenvalue problems for various special cases can be obtained. For example, if $\beta_0 = 0$, $\bar{\varepsilon}^{12} = \bar{\varepsilon}^{21} = \bar{\varepsilon}^{23} = \bar{\varepsilon}^{32} = 0$, and $\bar{\mu}^{12} = \bar{\mu}^{21} = \bar{\mu}^{23} = \bar{\mu}^{32} = 0$, then regardless of slant angle ζ only elements on the two diagonal lines of **M** are nonzero. So, the eigenvalue problem breaks up into two $2N \times 2N$ problems and the field components E_1 and H_2 are decoupled from E_2 and H_1. However, for slanted isotropic gratings in conical mounting the eigenvalue problem does not break up into two $2N \times 2N$ problems; the $2N \times 2N$ eigenvalue problems of Subsection 13.3.2.2 does not follow from the current formulation.

13.4 Crossed anisotropic gratings

For additional readings on the Fourier modal method for crossed gratings the reader may consult Chapter 7 of this book and references [13.22-28].

13.4.1 Description of the problem of crossed anisotropic gratings

A crossed grating can be defined as a planar layer of finite thickness whose optical properties vary periodically along two noncollinear directions. Other names are two-dimensional gratings and bigratings, but these names are sometimes used to mean other things. It is self evident that once a two dimensional structure is periodic in two noncollinear directions, it is periodic in countably infinite directions. We choose two directions that make an angle closest to $\pi/2$ as the principal periodic directions, or simply the periodic directions, of the grating, and define the smallest periods along the two directions as the periods of the grating. In most theoretical treatments of crossed gratings the two periodic directions are assumed to be mutually orthogonal, and many crossed gratings in applications indeed have this property. How-

ever, exceptions are common, so we do not make this assumption. The cost for accommodating nonorthogonality is not much, especially in dealing with anisotropic gratings.

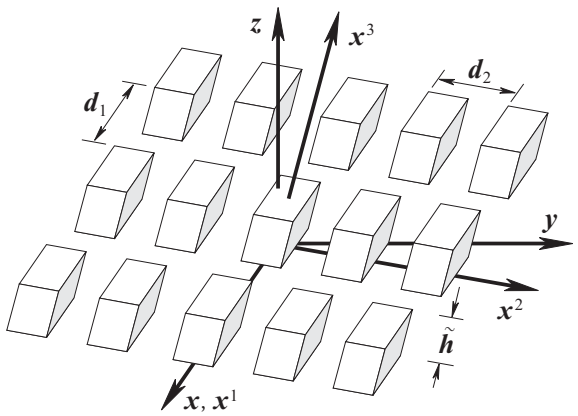


Fig. 13.6. Definition and notation of a crossed grating problem. All surfaces of the rhombs are parallel to a coordinate surface of the skew Cartesian coordinate system $Ox^1x^2x^3$ defined in Fig. 13.7.

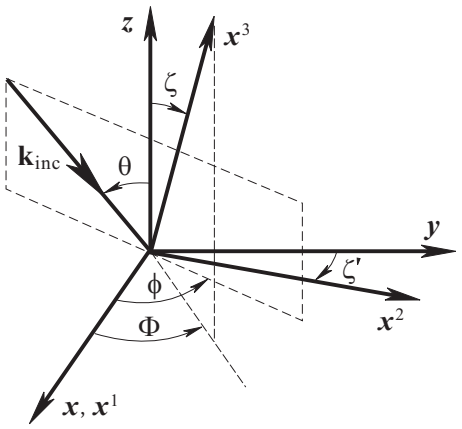


Fig. 13.7. Definition of the skew Cartesian coordinate system Ox1x2x3 relative to the rectangular Cartesian coordinate system Oxyz and definition of the angles of incidence. The axis x1 coincides with the axis x, and the three axes x1, x2, and y are all in the grating plane.

We consider the most general crossed gratings made with materials of anisotropic permittivity and permeability, but limit the grating profile to those that can be treated by the Fourier modal method the most efficiently, i.e., all grating surfaces are assumed to be parallel to one of the coordinate planes of a skew Cartesian coordinate system $Ox^1x^2x^3$. Figure 13.6 depicts such a crossed grating. The x^1 and x^2 axes are along the two periodic directions, so the x^1x^2 plane is the grating plane. The two grating periods are d_1 and d_2, and the pillar height $\tilde{h}$ is measured along the x^3 direction. The three axes are in general not mutually orthogonal and their relationship with respect to the conventional rectangular Cartesian coordinate system is defined in Fig. 13.7. The x axis coincides with the x^1 axis, and the xy plane coincides with the x^1x^2 plane. ζ' is the angle between the x^2 and y axes, ζ is the angle between the x^3 and z axes, and Φ is the angle between the x axis and the projection of the x^3 axis on the xy plane. Therefore, the two coordinate systems are related by

$$x = x^1 + x^2 \sin\zeta' + x^3 \sin\zeta \cos\Phi, \quad y = x^2 \cos\zeta' + x^3 \sin\zeta \sin\Phi, \quad z = x^3 \cos\zeta. \tag{13.110}$$

The covariant and contravariant basis vectors associated with this coordinate system are

$$\mathbf{b}_1 = \hat{\mathbf{x}}, \quad \mathbf{b}_2 = \hat{\mathbf{x}} \sin\zeta' + \hat{\mathbf{y}} \cos\zeta', \quad \mathbf{b}_3 = \hat{\mathbf{x}} \sin\zeta \cos\Phi + \hat{\mathbf{y}} \sin\zeta \sin\Phi + \hat{\mathbf{z}} \cos\zeta, \tag{13.111}$$

$$\begin{aligned} \mathbf{b}^1 &= \hat{\mathbf{x}} - \hat{\mathbf{y}} \tan\zeta' - \hat{\mathbf{z}} \tan\zeta \sec\zeta' \cos(\Phi + \zeta'), \\ \mathbf{b}^2 &= \hat{\mathbf{y}} \sec\zeta' - \hat{\mathbf{z}} \tan\zeta \sec\zeta' \sin\Phi, \\ \mathbf{b}^3 &= \hat{\mathbf{z}} \sec\zeta. \end{aligned} \tag{13.112}$$

respectively. The metric tensors $g^{\rho\sigma}$ and $g_{\rho\sigma}$ can be calculated from the basis vectors using formula (13.39). The permittivity tensor $\varepsilon^{\rho\sigma}$ and permeability tensor $\mu^{\rho\sigma}$ in this coordinate system are related to their counterparts $\overline{\varepsilon}^{\rho\sigma}$ and $\overline{\mu}^{\rho\sigma}$ in the rectangular Cartesian system in the same way as in (13.97).

For specification of incident angles and state of polarization of the incident plane wave we use the rectangular Cartesian coordinate system as usual. The incident polar angle θ and azimuth angle ϕ are defined in Fig. 13.7. The method of defining state of polarization is the same as in Subsection 13.3.1. The wave vector of the incident plane wave in the skew coordinate system is

$$\mathbf{k}_{\text{inc}} = \mathbf{b}^1 \alpha_0 + \mathbf{b}^2 \beta_0 + \mathbf{b}^3 \gamma_{00}^{(2)-}, \tag{13.113}$$

where

$$\begin{aligned} \alpha_0 &= k^{(2)} \sin\theta \cos\phi, \\ \beta_0 &= k^{(2)} \sin\theta \sin(\phi + \zeta'), \\ \gamma_{00}^{(2)-} &= k^{(2)} [\sin\zeta \sin\theta \cos(\Phi - \phi) - \cos\theta \cos\zeta]. \end{aligned} \tag{13.114}$$

The notation for the last covariant component will become apparent shortly.

13.4.2 Rayleigh expansions in skew three-dimensional coordinates

The electric field components in the two semi-infinite regions can be written in Rayleigh expansions:

$$\begin{aligned} E_\sigma^{(2)}(x^1, x^2, x^3) &= I_\sigma \exp[\mathrm{i}(\alpha_0 x^1 + \beta_0 x^2 + \gamma_{00}^{(2)-} x^3)] + \\ &\quad + \sum_{m,n} R_{\sigma mn} \exp[\mathrm{i}(\alpha_m x^1 + \beta_n x^2 + \gamma_{mn}^{(2)+} x^3)], \qquad x^3 \geq \tilde{h}; \end{aligned} \tag{13.115a}$$

$$E_\sigma^{(0)}(x^1, x^2, x^3) = \sum_{m,n} T_{\sigma mn} \exp[\mathrm{i}(\alpha_m x^1 + \beta_n x^2 + \gamma_{mn}^{(0)-} x^3)], \qquad x^3 \leq 0. \tag{13.115b}$$

where σ = 1, 2, 3, and I_σ, R_σ, and T_σ represent the incident, reflected, and transmitted electric field amplitudes, respectively. In the arguments of the exponential functions,

$$\alpha_m = \alpha_0 + m K_1, \qquad K_1 = 2\pi / d_1, \tag{13.116a}$$

$$\beta_n = \beta_0 + n K_2, \qquad K_2 = 2\pi / d_2, \tag{13.116b}$$

$$\gamma_{mn}^{(p)\pm} = [\pm k_{mn}^{3(p)} - (g^{31}\alpha_m + g^{32}\beta_n)] / g^{33}, \qquad p = 0, 2, \tag{13.117}$$

$$k_{mn}^{3(p)} = [g^{33}(k^{(p)2} - \alpha_m^2 g^{11} - 2\alpha_m \beta_n g^{12} - \beta_n^2 g^{22}) + (g^{31}\alpha_m + g^{32}\beta_n)^2]^{1/2}. \quad p = 0, 2. \tag{13.118}$$

From (13.115) we see that for a crossed grating it takes two indices to label a diffraction order. α_m, β_n, and $\gamma_{mn}^{(p)\pm}$ are the three covariant components and $k_{mn}^{3(p)}$ is the third contravariant component of the wave vector of the (m,n)th diffraction order, respectively. The sign of $k_{mn}^{3(p)}$ should be chosen so that

$$\mathrm{Re}[k_{mn}^{3(p)}] + \mathrm{Im}[k_{mn}^{3(p)}] > 0. \qquad p = 0, 2. \tag{13.119}$$

In this way the sum containing $\gamma_{mn}^{(p)+}$ in (13.115a) and that containing $\gamma_{mn}^{(p)-}$ in (115b) properly represent reflected and transmitted waves, respectively. The diffraction orders with a real $k_{mn}^{3(p)}$ are propagating and the rest with a nonzero imaginary part are evanescent.

The (m,n)th-order diffracted wave vector can be written as

$$\mathbf{k}_{mn}^{(p)} = \alpha_m \mathbf{b}^1 + \beta_n \mathbf{b}^2 + \gamma_{mn}^{(p)\pm} \mathbf{b}^3, \tag{13.120}$$

where the upper and lower signs are chosen for p = 2 and 0, respectively. The perpendicular projection of $\mathbf{k}_{mn}^{(p)}$ onto the grating plane, denoted by $\mathbf{k}_{mn}^{\perp}$, is the same in both regions:

$$\mathbf{k}_{mn}^{\perp} = \alpha_m \mathbf{b}_{\perp}^1 + \beta_n \mathbf{b}_{\perp}^2, \qquad \mathbf{b}_{\perp}^1 = \hat{\mathbf{x}} - \hat{\mathbf{y}} \tan\zeta', \qquad \mathbf{b}_{\perp}^2 = \hat{\mathbf{y}} \sec\zeta'. \tag{13.121}$$

$\mathbf{k}_{mn}^{\perp}$ can be written in another form

$$\mathbf{k}_{mn}^{\perp} = \mathbf{k}_{00}^{\perp} + \mathbf{K}_{mn}, \qquad \mathbf{K}_{mn} = m K_1 \mathbf{b}_{\perp}^1 + n K_2 \mathbf{b}_{\perp}^2. \tag{13.122}$$

Here $\mathbf{K}_{mn}$ defines the reciprocal space lattice of the crossed grating. It is reciprocal to the real space lattice

$$\mathbf{G}_{mn} = m d_1 \mathbf{b}_1 + n d_2 \mathbf{b}_2. \tag{13.123}$$

Figure 13.8 depicts an example of real space lattice and its reciprocal space lattice. Note that the periodic directions of the two lattices of differing indices are mutually orthogonal, but the periodic directions of the same indices are not parallel to each other, unless the lattices are rectangular. It is important to realize that $\mathbf{K}_{mn}$ together with $\mathbf{k}_{00}^{\perp}$, not $\mathbf{G}_{mn}$, gives the direction of the (m,n)th diffraction order in the grating plane.

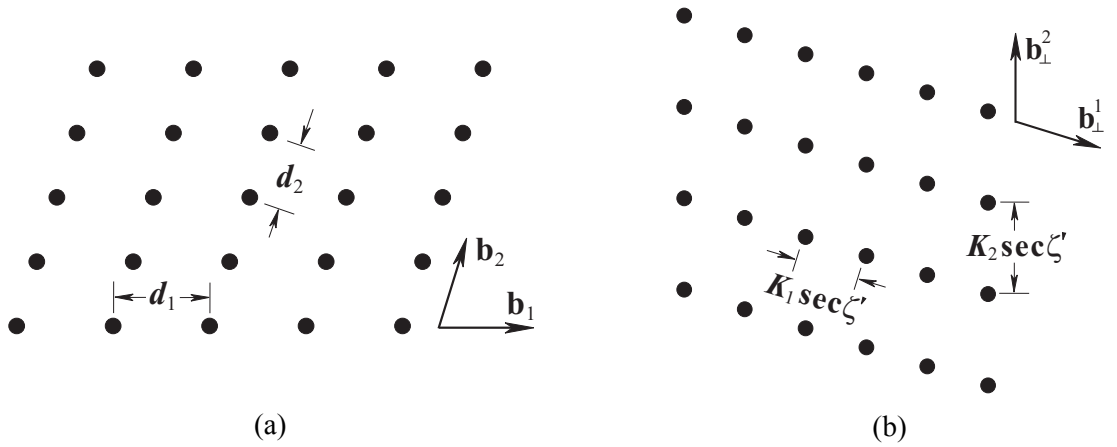


Fig. 13.8. Real space lattice (a) *and reciprocal space lattice* (b) of *a crossed grating.*

Among the six electromagnetic field components of a plane wave in a homogeneous medium, we choose E_1 and E_2 as the two independent ones. This is possible, unless $k_{mn}^{3(p)} = 0$ for some integers m and n. Since $k_{mn}^{3(p)}$ is the component of a diffracted order perpendicular to the grating plane, $k_{mn}^{3(p)} = 0$ means the diffracted order is propagating parallel to this plane, i.e., Rayleigh anomaly occurs. This situation can be avoided by slightly changing the incident angle, or wavelength, or grating period(s). The magnetic field components can be obtained from Maxwell equations. After this is done, the column vector of Fourier coefficients of the four total electromagnetic field components parallel to the grating plane in the two homogeneous regions are given by

$$\begin{pmatrix} E_{1mn}^{(p)} \\ E_{2mn}^{(p)} \\ H_{1mn}^{(p)} \\ H_{2mn}^{(p)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -C_{mn}^{(p)} & -A_{mn}^{(p)} & C_{mn}^{(p)} & A_{mn}^{(p)} \\ B_{mn}^{(p)} & C_{mn}^{(p)} & -B_{mn}^{(p)} & -C_{mn}^{(p)} \end{pmatrix} \begin{pmatrix} D_{mn}^{(p)+} & 0 & 0 & 0 \\ 0 & D_{mn}^{(p)+} & 0 & 0 \\ 0 & 0 & D_{mn}^{(p)-} & 0 \\ 0 & 0 & 0 & D_{mn}^{(p)-} \end{pmatrix} \begin{pmatrix} \tilde{u}_{1mn}^{(p)} \\ \tilde{u}_{2mn}^{(p)} \\ \tilde{d}_{1mn}^{(p)} \\ \tilde{d}_{2mn}^{(p)} \end{pmatrix}, \tag{13.124}$$

where $p = 0, 2$, $A_{mn}^{(p)}$, $B_{mn}^{(p)}$, $C_{mn}^{(p)}$, and $D_{mn}^{(p)\pm}$ are diagonal matrices with elements

$$A_{mn}^{(p)} = \frac{k^{(p)2} - \alpha_m^2}{k_0^{**} \mu^{(p)} k_{mn}^{3(p)}}, \quad B_{mn}^{(l)} = \frac{k^{(p)2} - \beta_n^2}{k_0^{**} \mu^{(p)} k_{mn}^{3(p)}}, \quad C_{mn}^{(p)} = \frac{\alpha_m \beta_n - \sin\zeta' k^{(p)2}}{k_0^{**} \mu^{(p)} k_{mn}^{3(p)}}, \tag{13.125a}$$

$$D_{mn}^{(p)+} = \exp[\mathrm{i}\gamma_{mn}^{(p)+}(x^3 - x_{p-1}^3)], \quad D_{mn}^{(p)-} = \exp[\mathrm{i}\gamma_{mn}^{(p)-}(x^3 - x_p^3)], \tag{13.125b}$$

$k_0^{**} = k_0 \cos\zeta \cos\zeta'$, $\mu^{(p)}$ is the scalar magnetic permeability of region p, $p = 0, 2$, $\tilde{u}_{\sigma mn}^{(2)} = R_{\sigma mn} \exp(\mathrm{i}\gamma_{mn}^{(2)+} x_1^3)$, $\tilde{u}_{\sigma mn}^{(0)} = 0$, $\tilde{d}_{\sigma mn}^{(2)} = I_\sigma \delta_{m0} \delta_{n0} \exp(\mathrm{i}\gamma_{00}^{(2)-} x_2^3)$, and $\tilde{d}_{\sigma mn}^{(0)} = T_{\sigma mn} \exp(\mathrm{i}\gamma_{mn}^{(0)-} x_0^3)$. The first and second square matrices in (13.124) are the $\mathrm{W}^{(p)}$ and $\phi^{(p)}$ matrices in the S matrix propagation algorithm. The diffraction efficiencies are given by

$$\eta_{mn}^{(2)} = A_{mn}^{(2)} |R_{2mn}|^2 + B_{mn}^{(2)} |R_{1mn}|^2 + C_{mn}^{(2)} (R_{1mn} \overline{R_{2mn}} + R_{2mn} \overline{R_{1mn}}), \tag{13.126a}$$

$$\eta_{mn}^{(0)} = A_{mn}^{(0)} |T_{2mn}|^2 + B_{mn}^{(0)} |T_{1mn}|^2 + C_{mn}^{(0)} (T_{1mn} \overline{T_{2mn}} + T_{2mn} \overline{T_{1mn}}), \tag{13.126b}$$

provided that

$$A_{00}^{(2)} |I_2|^2 + B_{00}^{(2)} |I_1|^2 + C_{00}^{(2)} (I_1 \overline{I_2} + I_2 \overline{I_1}) = 1, \tag{13.127}$$

and (m,n) makes $k_{mn}^{3(p)}$ real, where the bar means complex conjugation.

13.4.3 Fourier factorization of the constitutive relations

To apply the Fourier modal method to crossed anisotropic gratings we need to first write $\varepsilon^{\rho\sigma} E_\sigma$ and $\mu^{\rho\sigma} H_\sigma$ in two-dimensional Fourier space. Now $\varepsilon^{\rho\sigma}$ and $\mu^{\rho\sigma}$ are independent of x^3. Again, it is sufficient to work with $\varepsilon^{\rho\sigma} E_\sigma$ only. The Fourier transforms along x^1 and x^2 are to be done one at a time. Suppose we do it along x^1 first, then the analysis that we have carried out in Subsection 13.3.3.1 can be repeated here without change, because E_2, E_3, and D^1 are still continuous with respect to variable x^1, independent of parameter x^2. Therefore, from (13.99),

$$D_m^\rho(x^2) = (\varepsilon^{\rho\sigma} E_\sigma)_m(x^2) = \sum_n (\hat{\varepsilon}_1^{\rho\sigma})_{mn} E_{\sigma n}(x^2), \tag{13.128}$$

where we have used the subscript 1 in $\hat{\varepsilon}_1^{ij}$ to indicate that the Fourier transform is along x^1. Equation (13.128) is yet to be Fourier transformed along x^2, which can be done in a similar way. We write the equation as follows, dropping the x^2 dependence for simplicity,

$$D_m^1 = \sum_n \left\{ [\hat{\varepsilon}_1^{11} - \hat{\varepsilon}_1^{12}(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{21}]_{mn} E_{1n} + [\hat{\varepsilon}_1^{12}(\hat{\varepsilon}_1^{22})^{-1}]_{mn} D_n^2 + [\hat{\varepsilon}_1^{13} - \hat{\varepsilon}_1^{12}(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{23}]_{mn} E_{3n} \right\}, \tag{13.129a}$$

$$E_{2m} = \sum_n \left\{ (\hat{\varepsilon}_1^{22})_{mn}^{-1} D_n^2 - [(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{21}]_{mn} E_{1n} - [(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{23}]_{mn} E_{3n} \right\}, \tag{13.129b}$$

$$D_m^3 = \sum_n \left\{ [\hat{\varepsilon}_1^{31} - \hat{\varepsilon}_1^{32}(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{21}]_{mn} E_{1n} + [\hat{\varepsilon}_1^{32}(\hat{\varepsilon}_1^{22})^{-1}]_{mn} D_n^2 + [\hat{\varepsilon}_1^{33} - \hat{\varepsilon}_1^{32}(\hat{\varepsilon}_1^{22})^{-1}\hat{\varepsilon}_1^{23}]_{mn} E_{3n} \right\}, \tag{13.129c}$$

On the right-hand sides of (13.129) every one of the Fourier coefficient E_{1n}, E_{3n}, and D_n^2 is a continuous function of x^2 (for a proof, see [13.25]); therefore, Laurent's rule can be applied to each scalar product. This gives the result

$$D_{mn}^{\rho} = (\varepsilon^{\rho\sigma} E_{\sigma})_{mn} = \sum_{i,j} (\hat{\varepsilon}_{12}^{\rho\sigma})_{mn,ij} E_{\sigma ij}, \tag{13.130}$$

where

$$\hat{\boldsymbol{\varepsilon}}_{12} = l_2^+ F_2 l_2^- (\hat{\boldsymbol{\varepsilon}}_1) = l_2^+ F_2 l_2^- \, l_1^+ F_1 l_1^- (\boldsymbol{\varepsilon}), \tag{13.131}$$

and we have used the second subscript in $\hat{\varepsilon}_{12}^{\rho\sigma}$ to indicate that the second-time Fourier transform is along x^2.

Unlike the one-dimensional Fourier factorization of $\varepsilon^{\rho\sigma} E_{\sigma}$ where the result is unique, in two-dimensional Fourier factorization the result is not unique. In the above, we could have reversed the order of Fourier transforms, i.e., first along x^2 then along x^1. This would lead to

$$\hat{\boldsymbol{\varepsilon}}_{21} = l_1^+ F_1 l_1^- (\hat{\boldsymbol{\varepsilon}}_2) = l_1^+ F_1 l_1^- \, l_2^+ F_2 l_2^- (\boldsymbol{\varepsilon}), \tag{13.132}$$

which is an equally acceptable expression. In fact, there are a multitude of choices. For example, we can use $\hat{\varepsilon}_{12}^{\rho\sigma}$ for some ρσ combinations and $\hat{\varepsilon}_{21}^{\rho\sigma}$ for the rest, or simply take average of $\hat{\varepsilon}_{12}^{\rho\sigma}$ and $\hat{\varepsilon}_{21}^{\rho\sigma}$ for all ρ and σ. Presumably all choices lead to the same numerical result in the limit of infinite un-truncated Fourier matrices. In finite truncated Fourier space they are slightly different. In most practical applications the differences are unimportant when the numerical results have converged, but under some very special circumstances where the incident plane wave condition, the grating medium anisotropy, and the grating profile altogether support certain symmetry, a choice compatible with the symmetry is obviously preferred. An example of such a case can be found in [13.7]. Ignoring the slight differences, we will drop the subscripts and use $\hat{\varepsilon}^{\rho\sigma}$ to denote the two-dimensional Fourier representation of $\boldsymbol{\varepsilon}$ in what follows.

13.4.4 Fields in the two-dimensionally periodic anisotropic region

Once the Fourier representations of the constitutive relations are obtained, the derivation of the modal fields and the total fields is straight-forward. A covariant component of any modal field vector takes this form

$$F_{\sigma}(x^1, x^2, x^3) = \exp(\mathrm{i}\gamma x^3) \sum_{m,n} \exp(\mathrm{i}\alpha_m x^1 + \mathrm{i}\beta_n x^2) F_{\sigma mn}, \tag{13.133}$$

where F stands for E or H, and σ = 1, 2, 3. After examining the derivation in Subsection 13.3.3.2 the reader can see that almost all results there can be copied here. In particular, (13.103) through (13.107) are valid for crossed anisotropic gratings, provided that k_0^* is re-

placed by k_0^{**}, the scalar β_0 is replaced by diagonal matrix $\boldsymbol{\beta}$ whose elements are $\delta_{mi}\delta_{nj}\beta_n$, and $\boldsymbol{\alpha}$ is understood as $\delta_{mi}\delta_{nj}\alpha_m$. For the sake of saving space, we will not repeat all equations here, except to give the matrix of the eigenvalue problem and the expression of the total fields:

$$\mathbf{M}=\begin{pmatrix} -\tilde{\mu}^{23}\boldsymbol{\beta}-\boldsymbol{\alpha}\tilde{\varepsilon}^{31} & \tilde{\mu}^{23}\boldsymbol{\alpha}-\boldsymbol{\alpha}\tilde{\varepsilon}^{32} & k_0^{**}\tilde{\mu}^{21}+\dfrac{1}{k_0^{**}}\boldsymbol{\alpha}\tilde{\varepsilon}^{33}\boldsymbol{\beta} & k_0^{**}\tilde{\mu}^{22}-\dfrac{1}{k_0^{**}}\boldsymbol{\alpha}\tilde{\varepsilon}^{33}\boldsymbol{\alpha} \\ \tilde{\mu}^{13}\boldsymbol{\beta}-\boldsymbol{\beta}\tilde{\varepsilon}^{31} & -\tilde{\mu}^{13}\boldsymbol{\alpha}-\boldsymbol{\beta}\tilde{\varepsilon}^{32} & -k_0^{**}\tilde{\mu}^{11}+\dfrac{1}{k_0^{**}}\boldsymbol{\beta}\tilde{\varepsilon}^{33}\boldsymbol{\beta} & -k_0^{**}\tilde{\mu}^{12}-\dfrac{1}{k_0^{**}}\boldsymbol{\beta}\tilde{\varepsilon}^{33}\boldsymbol{\alpha} \\ -k_0^{**}\tilde{\varepsilon}^{21}-\dfrac{1}{k_0^{**}}\boldsymbol{\alpha}\tilde{\mu}^{33}\boldsymbol{\beta} & -k_0^{**}\tilde{\varepsilon}^{22}+\dfrac{1}{k_0^{**}}\boldsymbol{\alpha}\tilde{\mu}^{33}\boldsymbol{\alpha} & -\tilde{\varepsilon}^{23}\boldsymbol{\beta}-\boldsymbol{\alpha}\tilde{\mu}^{31} & \tilde{\varepsilon}^{23}\boldsymbol{\alpha}-\boldsymbol{\alpha}\tilde{\mu}^{32} \\ k_0^{**}\tilde{\varepsilon}^{11}-\dfrac{1}{k_0^{**}}\boldsymbol{\beta}\tilde{\mu}^{33}\boldsymbol{\beta} & k_0^{**}\tilde{\varepsilon}^{12}+\dfrac{1}{k_0^{**}}\boldsymbol{\beta}\tilde{\mu}^{33}\boldsymbol{\alpha} & \tilde{\varepsilon}^{13}\boldsymbol{\beta}-\boldsymbol{\beta}\tilde{\mu}^{31} & -\tilde{\varepsilon}^{13}\boldsymbol{\alpha}-\boldsymbol{\beta}\tilde{\mu}^{32} \end{pmatrix}, \quad (13.134)$$

$$\begin{pmatrix} E_1 \\ E_2 \\ H_1 \\ H_2 \end{pmatrix}=\sum_{m,n=m_1}^{m_2}\exp(\mathrm{i}\alpha_m x^1+\mathrm{i}\beta_n x^2)\sum_{q=1}^{2N^2}\begin{pmatrix} E^+_{1mnq} & E^-_{1mnq} \\ E^+_{2mnq} & E^-_{2mnq} \\ H^+_{1mnq} & H^-_{1mnq} \\ H^+_{2mnq} & H^-_{2mnq} \end{pmatrix}\begin{pmatrix} \exp[\mathrm{i}\gamma_q^+(x^3-x_0^3)] & 0 \\ 0 & \exp[\mathrm{i}\gamma_q^-(x^3-x_1^3)] \end{pmatrix}\begin{pmatrix} \tilde{u}_q \\ \tilde{d}_q \end{pmatrix}. \quad (13.135)$$

In (13.134) $\tilde{\boldsymbol{\varepsilon}}=l_3^-(\hat{\boldsymbol{\varepsilon}})$ and $\tilde{\boldsymbol{\mu}}=l_3^-(\hat{\boldsymbol{\mu}})$, and in (13.135) we have assumed that the Fourier space truncation in both x^1 and x^2 directions are from m_1 to $m_2 = N + m_1 - 1$. The eigenvalue partition follows the same principle as explained previously. The eigenvector matrix in (13.135) gives the $\mathbf{W}^{(1)}$ matrix needed to carry out the S-matrix propagation algorithm.

13.5 Staircase approximation and S-matrix algorithm

In this chapter by assuming invariance of medium permittivity and permeability of the grating layer in an out-of-plane direction, the z or x^3 direction, we are able to turn the problem of solving Maxwell equations into that of finding solutions of an eigenvalue problem. The Fourier modal method is the most suitable for this type of grating structures. In particular, its degree of difficulty in numerical solution is independent on groove depth or pillar height. However, the assumption also restricts the applicability of the Fourier modal method because the structural invariance is not always available.

13.5.1 Staircase approximation

A common method to extend the Fourier modal method to grating structures varying in the x^3-direction within the grating layer is to apply the staircase approximation. A grating of arbitrary x^3-direction variation is sliced into many layers parallel to the grating plane and in each layer the medium boundary is locally replaced by an x^3-invariant boundary. Then within each modified layer the Fourier modal method is applicable. The total fields in a modified layer are connected to those of neighboring layers by electromagnetic boundary conditions at the layer interfaces, and ultimately to the two semi-infinite homogeneous regions. In this way solution to the original grating problem is "approximated" by solution to the modified grating problem.

From an intuitive and naïve point of view the staircase approximation seems reasonable. As the number of layers tends to infinity and the maximum layer thickness tends to zero, the modified structure tends the original one, if the observation is made on a fixed length scale. This idea of approximation is similar to that used in elementary calculus to calculate the area

of an arbitrary shape by integrating rectangular infinitesimal area elements d*x*d*y*. However, the problem here is physical, not mathematical. Ample numerical experiments have shown that, in case of one-dimensionally periodic gratings, the approximation works well for TE polarization, but not well for TM polarization especially for highly conducting metallic gratings. The reason is electromagnetic. At the sharp edge of a wedge formed by nonmagnetic media of different permittivities, the electric field component parallel to the edge direction is finite but that transverse to the edge direction, if nonzero, is infinite [13.29]. In staircase approximation many edges are artificially created. In TE polarization the electric field remains finite everywhere, so no severe numerical problem is introduced. In TM polarization, the electric field that should be finite near the smooth grating profile becomes infinitely large at the edges of the staircase boundary. In other word, the near field is altered completely. Therefore, the numerical convergence and computation accuracy of far-field properties, such as diffraction efficiency, is severely affected. For one-dimensional gratings in conical mounting and crossed gratings under any incidence condition, electric field components transverse to the sharp edges are unavoidable when staircase approximation is used.

The artificially introduced adverse effect of staircase approximation and its cause were known for a long time, but the first detailed study appeared much later [13.30]. If the approximation were used to study properties of a periodic structure of negative permittivity free of sharp edges, something much worse can happen. For certain combination of negative permittivity and wedge angle no numerical method converges to date [13.31-33].

From computation efficiency point of view staircase approximation is also inadvisable. The computation time required by the approximation is roughly proportional to the number of layers. The more layers are used, the more accurate the numerical results are obtained, but also the more computation time is needed. In general, when a grating of non-staircase profile is encountered, it is advisable to use a method that does not rely on staircase approximation, such as the integral method described in Chapter 4, or the differential method in Chapter 7, or the C method in Chapter 8.

The advantage of staircase approximation, if it works, is that it is simple to implement algorithmically. So, one has to make a tradeoff between programming complexity and computation efficiency. The best way to implement the staircase approximation is to use the S-matrix propagation algorithm.

13.5.2 S-matrix algorithm

Suppose the space is divided into $P + 1$ layers in the direction normal to the grating plane, where $P \geq 1$. Layer 0 is the semi-infinite substrate and layer $P + 1$ is the semi-infinite cover. Layers 1 through layer P are the staircase approximation layers. We assume medium p is between $x^3 = x_{p-1}^3$ and $x^3 = x_p^3$ and has thickness $\tilde{h}_p = x_p^3 - x_{p-1}^3$. As demonstrated repeatedly in this chapter the Fourier coefficients of the vector components of the total tangential fields can be written in a standard form:

$$\mathscr{F}^{(p)}(x^3) = (W^{(p)+},\, W^{(p)-}) \begin{pmatrix} \exp[\mathrm{i}\gamma^{(p)+}(x^3 - x_{p-1}^3)] & 0 \\ 0 & \exp[\mathrm{i}\gamma^{(p)-}(x^3 - x_p^3)] \end{pmatrix} \begin{pmatrix} \tilde{u}^{(p)} \\ \tilde{d}^{(p)} \end{pmatrix}, \quad (13.136)$$

with

$$\tilde{u}_q^{(p)} = u_q^{(p)} \exp(\mathrm{i}\gamma_q^{(p)+} x_{p-1}^3), \qquad \tilde{d}_q^{(p)} = d_q^{(p)} \exp(\mathrm{i}\gamma_q^{(p)-} x_p^3), \quad (13.137)$$

where $W^{(p)+}$ and $W^{(p)-}$ are 2×1 block matrices and for simplicity we have omitted the subscripts for eigenvalues and modal amplitudes. Matching boundary conditions at $x^3 = x_p^3$ gives $\mathscr{F}^{(p)}(x_p^3 - 0) = \mathscr{F}^{(p+1)}(x_p^3 + 0)$, i.e.,

$$(W^{(p+1)+},\ W^{(p+1)-})\begin{pmatrix}1 & 0\\ 0 & \phi_-^{(p+1)-1}\end{pmatrix}\begin{pmatrix}\tilde{u}^{(p+1)}\\ \tilde{d}^{(p+1)}\end{pmatrix}=(W^{(p)+},\ W^{(p)-})\begin{pmatrix}\phi_+^{(p)} & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tilde{u}^{(p)}\\ \tilde{d}^{(p)}\end{pmatrix}, \tag{13.138}$$

where

$$\phi_-^{(p)-1}=\exp(-\mathrm{i}\gamma^{(p)-}\tilde{h}_p),\qquad \phi_+^{(p)}=\exp(\mathrm{i}\gamma^{(p)+}\tilde{h}_p). \tag{13.139}$$

In the S-matrix algorithm we seek a set of matrices $S^{(p)}$ that links the inputs and outputs of the layer assemble 0 through p, such that

$$\begin{pmatrix}\tilde{u}^{(p)}\\ \tilde{d}^{(0)}\end{pmatrix}=S^{(p-1)}\begin{pmatrix}\tilde{u}^{(0)}\\ \tilde{d}^{(p)}\end{pmatrix}=\begin{pmatrix}T_{\mathrm{uu}}^{(p-1)} & R_{\mathrm{ud}}^{(p-1)}\\ R_{\mathrm{du}}^{(p-1)} & T_{\mathrm{dd}}^{(p-1)}\end{pmatrix}\begin{pmatrix}\tilde{u}^{(0)}\\ \tilde{d}^{(p)}\end{pmatrix}. \tag{13.140}$$

From (13.138) and (13.140) we obtain a set of recursion formulas for getting $S^{(p)}$ from $S^{(p-1)}$ and the **W** matrices directly:

$$T_{\mathrm{uu}}^{(p)}=(\mathbf{Z}^{-1}X_1)_1, \tag{13.141a}$$

$$R_{\mathrm{ud}}^{(p)}=(\mathbf{Z}^{-1}X_2)_1, \tag{13.141b}$$

$$R_{\mathrm{du}}^{(p)}=R_{\mathrm{du}}^{(p-1)}+T_{\mathrm{dd}}^{(p-1)}(\mathbf{Z}^{-1}X_1)_2, \tag{13.141c}$$

$$T_{\mathrm{dd}}^{(p)}=T_{\mathrm{dd}}^{(p-1)}(\mathbf{Z}^{-1}X_2)_2, \tag{13.141d}$$

where

$$\begin{aligned}\mathbf{Z}&=(W^{(p+1)+},\ -W^{(p)+}\phi_+^{(p)}R_{\mathrm{ud}}^{(p-1)}-W^{(p)-}),\\ X_1&=W^{(p)+}\phi_+^{(p)}T_{\mathrm{uu}}^{(p-1)},\qquad X_2=-W^{(p+1)-}\phi_-^{(p+1)-1}.\end{aligned} \tag{13.142}$$

In (13.141) the subscripts 1 and 2 attached to the parentheses refer to the upper and lower block of the matrix enclosed. The S-matrix recursion can be initialized with $S^{(-1)}$ being an identity matrix, and continued until $S^{(P)}$ is obtained, which links the Rayleigh amplitudes in the two semi-infinite regions. For most applications $\tilde{u}^{(0)}=0$, so only recursions of $R_{\mathrm{ud}}^{(p)}$ and $T_{\mathrm{dd}}^{(p)}$ are necessary. Furthermore, if only the reflected diffraction orders are of interest, only recursion of $R_{\mathrm{ud}}^{(p)}$ is necessary.

Note that both exponential functions in (13.139) decay to zero as the layer thickness $\tilde{h}_p$ increases or as the absolute values of the imaginary parts of the eigenvalues increase, thanks to our proper eigenvalue partition. Furthermore, the two exponential functions are not inverted in (13.142). This is the key to the numerical stability of the S-matrix algorithm.

13.6 Concluding Remarks

In this chapter we have presented the basic structure and ingredients of the Fourier modal method. Like most other methods the fields in the homogeneous regions are expanded in Rayleigh expansions. The core idea of the method, as its name says and as Subsections 13.2.1-13.2.3 illustrate, is to solve the modal fields in the grating layer in discrete Fourier space. All other aspects, such as the treatments of slanted grating profile, conical mounting, anisotropic medium, two-dimensional periodicity (crossed grating), are technical details. The modal approach is permitted by the invariance of the grating layer along an out-of-plane direction.

After 50 years of development, the Fourier modal method has become one of the most popular methods, if not the most popular method, for modeling diffraction gratings. However,

it will be wrong to say that it is the best method or it can solve all grating problems. The first statement is wrong because each method has its strength and weakness and without knowing the specifics of a concrete grating problem it is impossible to assess which method is the best for the given problem. The second statement is wrong as we have seen a counterexample in Subsection 13.5.1. In general the Fourier modal method is most suitable to treat gratings with the out-of-plane direction invariance, and its most prominent weakness is its inefficiency or inability to handle metallic gratings without this invariance in TM polarization.

We have assumed that the whole grating surface consists of facets and every facet coincides with a coordinate surface of a suitably chosen skew Cartesian coordinate system. The x^3-invariance is dictated by the modal approach and the top and bottom facets being coincided with two constant x^3 surfaces is required to accommodate the staircase approximation. For the crossed grating problem, the requirement that the other grating facets be parallel to coordinate surfaces of constant x^1 and x^2 is made only to simplify Fourier factorization of the constitutive relations. In real crossed gratings this requirement is often not met and cannot be viewed as a reasonable approximation. In the original work of [13.24] a boundary of arbitrary constant x^3 cross section was approximated by a zigzag boundary, so Fourier factorization can be easily made. This approximation is equivalent to the staircase approximation in the vertical direction, and therefore suffers from the same problem as commented in Subsection 13.5.1. Popov and Nevière [13.34] later found a way to make Fourier factorization without the zigzag approximation. This important contribution to the Fourier factorization theory was first termed fast Fourier factorization, but now the more acceptable term seems to be normal vector field method as proposed by Schuster *et al.* [13.26]. While the x^3-invariance is fundamentally important to the Fourier modal method, there is no reason that the in-plane coordinate system has to be Cartesian. Recently Weiss *et al*. developed the matched-coordinate Fourier modal method [13.7], in which the general x^3-invariant curvilinear coordinate surfaces match the sidewall surface of the crossed gratings. Convergence faster than that of the normal-vector-field method has been achieved. Both the works on the normal-vector-field method and the matched-coordinate method further extend the Fourier modal method and their developments can be fit in the general framework of Section 13.4.

The Fourier modal method relies on the periodicity of the gratings; it is the periodicity that makes discrete Fourier expansion possible. It seems natural to use Fourier basis for analysis of periodic structures; however, Fourier basis has its drawback in handling discontinuous periodic functions because its basis functions are continuous functions. Since mid 1990s and especially in recent years some interesting research works have been done to expand the modal fields in terms of discontinuous basis functions or to use orthogonal polynomial expansions separately for each constant permittivity section of a grating period [13.35-40]. Nonetheless, to date the Fourier basis remains to be the most convenient basis to use.

References

1. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," J. Opt. Soc. Am. **71**, 811-818 (1981).
2. H. Kogelnik, "Coupled wave theory for thick hologram gratings," Bell Syst. Tech. J. **48**, 2909-2947 (1969).
3. M. G. Moharam, E. B. Grann, D. A. Pommet, and T. K. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," J. Opt. Soc. Am. A **12**, 1068-1076 (1995).
4. M. G. Moharam, D. A. Pommet, E. B. Grann, and T. K. Gaylord, "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmission matrix approach," J. Opt. Soc. Am. A **12**, 1077-1086 (1995).
5. L. Li, "Mathematical reflections on the Fourier modal method in grating theory," in *Mathematical Modeling in Optical Science*, SIAM (Society for Industrial and Applied Mathematics) Frontiers in Applied Mathematics (SIAM, Philadelphia, 2001), Eds. G. Bao, L. Cowsar, and W. Masters, pp. 111-139.
6. G. Granet, "Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution," J. Opt. Soc. Am. A **16**, 2510-2516 (1999).
7. T. Weiss, G. Granet, N. A. Gippius, S. G. Tikhodeev, and H. Giessen, "Matched coordinates and adaptive spatial resolution in the Fourier modal method," Opt. Express **17**, 8051-8061 (2009).
8. L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," J. Opt. Soc. Am. A **13**, 1870-1876 (1996).
9. L. Li, "Justification of matrix truncation in the modal methods of diffraction gratings," J. Opt. A: Pure Appl. Opt. **1**, 531-536 (1999).
10. Ph. Lalanne and G. M. Morris, "Highly improved convergence of the coupled-wave method for TM polarization," J. Opt. Soc. Am. A **13**, 779-784 (1996).
11. G. Granet and B. Guizal, "Efficient implementation of the coupled-wave method for metallic lamellar gratings in TM polarization," J. Opt. Soc. Am. A **13**, 1019-1023 (1996).
12. L. Li, "Note on the S-matrix propagation algorithm," J. Opt. Soc. Am. A **20**, 655-660 (2003).
13. L. Li, "Formulation and comparison of two recursive matrix algorithms for modeling layered diffraction gratings," J. Opt. Soc. Am. A **13**, 1024-1035 (1996).
14. E. L. Tan, "Note on formulation of the enhanced scattering- (transmittance-) matrix approach," J. Opt. Soc. Am. A **19**, 1157-1161 (2002).
15. M. G. Moharam and T. K. Gaylord, "Three-dimensional vector coupled-wave analysis of planar grating diffraction," J. Opt. Soc. Am. 7**3**, 1105-1112 (1983).
16. B. Chernov, N. Nevière, and E. Popov, "Fast Fourier factorization method applied to modal method analysis of slanted lamellar diffraction gratings in conical mountings," Opt. Commun. **194**, 289-297 (2001).
17. K. Rokushima and J. Yamakita, "Analysis of anisotropic dielectric gratings," J. Opt. Soc. Am. **73**, 901-908 (1983).

18. E. N. Glytsis and T. K. Gaylord, "Rigorous three-dimensional coupled-wave diffraction analysis of single and cascaded anisotropic gratings," J. Opt. Soc. Am. A **4**, 2061-2080 (1987).

19. S. Mori, K. Mukai, and J. Yamakita, "Analysis of dielectric lamellar gratings coated with anisotropic layers," J. Opt. Soc. Am. A **7**, 1661-1665 (1990).

20. L. Li, "Reformulation of the Fourier modal method for surface-relief gratings made with anisotropic materials," J. Mod. Opt. **45**, 1313-1334 (1998).

21. L. Li, "Oblique-coordinate-system-based Chandezon method for modeling one-dimensionally periodic, multilayer, inhomogeneous, anisotropic gratings," J. Opt. Soc. Am. A, **16**, 2521-2531 (1999).

22. R. Bräuer and O. Bryngdahl, "Electromagnetic diffraction analysis of two-dimensional gratings," Opt. Commun. **100**, 1-5 (1993).

23. E. Noponen and J. Turunen, "Eigenmode method for electromagnetic synthesis of diffractive elements with three-dimensional profiles," J. Opt. Soc. Am. A **11**, 2494-(1994).

24. L. Li, "New formulation of the Fourier modal method for crossed surface-relief gratings," J. Opt. Soc. Am. A **14**, 2758-2767 (1997).

25. L. Li, "Fourier modal method for crossed anisotropic gratings with arbitrary permittivity and permeability tensors," J. Opt. A: Pure Appl. Opt. **5**, 345-355 (2003).

26. T. Schuster, J. Ruoff, N. Kerwien, S. Rafler, and W. Osten, "Normal vector method for convergence improvement using the RCWA for crossed gratings," J. Opt. Soc. Am. A **24**, 2880-2890 (2007).

27. R. Antos, "Fourier factorization with complex polarization bases in modeling optics of discontinuous bi-periodic structures," Opt. Express **17**, 7269-7274 (2009).

28. M. Onishi, K. Crabtree, and R. A. Chipman, "Formulation of rigorous coupled-wave theory for gratings in bianisotropic media," J. Opt. Soc. Am. A, **28**, 1758-(2011).

29. J. van Bladel, *Singular Electromagnetic Fields and Sources* (Clarendon, Oxford, 1991).

30. E. Popov, M. M. Nevière, B. Gralak, and G. Tayeb, "Staircase approximation validity for arbitrary-shaped gratings," J. Opt. Soc. Am. A **19**, 33-42 (2002).

31. L. Li and G. Granet, "Field singularities at lossless metal dielectric right-angle edges and their ramifications to the numerical modeling of gratings," J. Opt. Soc. Am. A **28**, 738-746 (2011).

32. L. Li, "Field singularities at lossless metal dielectric arbitrary-angle edges and their ramifications to the numerical modeling of gratings," J. Opt. Soc. Am. A **29**, 593-604 (2012).

33. L. Li, "Hypersingularity, electromagnetic edge condition, and an analytic hyperbolic wedge model," J. Opt. Soc. Am. A **31** (accepted for publication, Feb. 2014).

34. E. Popov and M. Nevière, "Maxwell equations in Fourier space: fast-converging formulation for diffraction by arbitrary shaped, periodic, anisotropic media," J. Opt. Soc. Am. A **18**, 2886-2894 (2001).

35. R. H. Morf, "Exponentially convergent and numerically efficient solution of Maxwell's equations for lamellar gratings," J. Opt. Soc. Am. A **12**, 1043-1056 (1995).

36. P. Lalanne and J.-P. Hugonin, "Numerical performance of finite-difference modal methods for the electromagnetic analysis of one-dimensional lamellar gratings," J. Opt. Soc. Am. A **17**, 1033-1042 (2000).

37. P. Bouchon, F. Pardo, R. Haïdar, and J.-L. Pelouard, "Fast modal method for subwavelength gratings based on B-spline formulation," J. Opt. Soc. Am. A **27**, 696-702 (2010).

38. G. Granet, L. B. Andriamanampisoa, K. Raniriharinosy, A. M. Armeanu, and K. Edee, "Modal analysis of lamellar gratings using the moment method with subsectional basis and adaptive spatial resolution," J. Opt. Soc. Am. A **27**, 1303-1310 (2010).

39. D. Song, L. Yuan, and Y. Y. Lu, "Fourier-matching pseudospectral modal method for diffraction gratings," J. Opt. Soc. Am. A **28**, 613-620 (2011).

40. K. Edee, "Modal method based on subsectional Gegenbauer polynomial expansion for lamellar gratings," J. Opt. Soc. Am. A **28**, 2006-2013 (2011).